Using Geovisual Analytics to investigate the performance of Geographically Weighted Discriminant Analysis

Author : Peter Foley

Supervisor : Dr. Urška Demšar

Head of Department : Prof. Rob Kitchin

Submission Date : February 2012

Thesis submitted in fulfillment of the requirements of the M.Sc. by Research Degree

National Centre for Geocomputation

Faculty of Science

National University of Ireland Maynooth

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Urška Demšar for her advice and assistance. I could not have completed this research without her guidance.

I would also like to thank Prof. Chris Brunsdon from the University of Liverpool, UK; Dr. Frank Hardisty and Dr. Anthony Robinson from the GeoVISTA Center at Pennsylvania State University, USA; Prof. Stewart Fotheringham from the University of St. Andrews, UK; Mr. Martin Charlton from the National Centre for Geocomputation at the National University of Ireland, Maynooth, Ireland; Prof. Jason Dykes and Dr. Jo Wood from the giCentre at City University, UK for their suggestions and advice.

Finally, I would like to thank Science Foundation Ireland who have funded this research through a Research Frontiers Programme Grant (09/RFP/CMS2250) awarded to Urška Demšar under the National Development Plan.

Abstract

Geographically Weighted Discriminant Analysis (GWDA) is a method for prediction and analysis of categorical spatial data. It is an extension of Linear Discriminant Analysis (LDA) that allows the relationship between the predictor variables and the categories to vary spatially. This is also referred to spatial non-stationarity. If spatial non-stationarity exists, GWDA should model the relationship between the categories and predictor variables more accurately, thus resulting in a lower classification uncertainty and ultimately a higher classification accuracy. The GWDA output also requires interpretation to understand which variables are important in driving the classification in different geographical regions. This research uses interactive visualisations from the field of geovisual analytics to investigate the performance of GWDA in terms of classification accuracy, classification uncertainty and spatial non-stationarity. The methodology is demonstrated in a case study that uses GWDA to examine the relationship between county level voting patterns in the 2004 US presidential election and five socio-economic indicators. This research builds on existing techniques to interpret the GWDA output and provides additional insight into the processes driving the classification. It also demonstrates a practical application of geovisual analytic tools.

Contents

List of Tables

List of Figures

List of Acronyms

CIE Commission Internationale de l'Eclairage

ESTAT Exploratory Spatio-Temporal Analysis Toolkit

GWDA Geographically Weighted Discriminant Analysis

GWPCA Geographically Weighted Principal Components Analysis

GWR Geographically Weighted Regression

KDA Kernel Discriminant Analysis

LDA Linear Discriminant Analysis

LOWESS Locally Weighted Scatterplot Smoothing

NVAC National Visualization and Analytics Center

OSL Ordering Single Link

PCA Principal Components Analysis

PCP Parallel Coordinates Plot

PDF Probability Density Function

QDA Quadratic Discriminant Analysis

STGWR Spatio-Temporal Geographically Weighted Regression

US United States of America

UK United Kingdom

1 Introduction

Geographically Weighted Discriminant Analysis (GWDA) is a spatial statistical method that models a spatially varying relationship between a categorical dependent variable and a set of continuous predictor variables. The output from GWDA is complex and multivariate in nature and therefore, difficult to interpret analytically. This thesis uses techniques from the field of geovisual analytics to help interpret the GWDA output.

1.1 Spatial statistics

Statistical modelling and analysis of spatial data require a special approach because of the unique characteristics of spatial data. These characteristics include: the modifiable areal unit problem, boundary problems, spatial sampling procedures and spatial autocorrelation (Rogerson 2008). Spatial autocorrelation or spatial dependence means that geographically closer objects tend to be more similar than distant objects and this violates the principle of independence which is an assumption of classical statistics (Reimann et al. 2008). Additionally, common statistical modelling techniques such as regression typically assume that modelled relationships are constant over space. Although this is a reasonable assumption in the physical sciences, it is not necessarily true in the social sciences where "some relationships are intrinsically different across space" (Fotheringham et al. 2002). The concept of spatially varying relationships is referred to as spatial non-stationarity, or spatial heterogeneity (Crespo 2009). The field of spatial statistics provides a set of methods to deal with spatial autocorrelation and spatial non-stationarity in spatial data.

Relationships in non-spatial data are typically modelled by fitting a particular mathematical function to the data. For example, regression models the relationship between a set of predictor variables and a continuous dependent variable as a line or other simple function. However, this approach breaks down if the relationship between the dependent variable and the predictor variables is sufficiently complex and cannot be described analytically. One solution is to fit simple mathematical functions to local subsets of the data. This allows a complex relationship to be decomposed and estimated as a set of continuous polynomials. Spline regression or Locally Weighted Scatterplot Smoothing (LOWESS) curves (Afifi et al. 2004) are examples of local regression models. The fitting of distinct models to local subsets of the data is referred to as 'local' modelling to distinguish it from a single 'global' model fitted to the entire data set.

The local modelling approach can also be adapted to deal with spatial autocorrelation and spatial non-stationarity in spatial data. In the case of geographically local models, each geographical location is characterised by a distinct model. The expansion method and Geographically Weighted Regression (GWR) are examples of local spatial regression models used to model spatial non-stationarity (Rogerson 2008).

GWR (Fotheringham et al. 2002) models spatial non-stationarity by allowing the parameters of a linear regression function to vary spatially. The GWR model is calibrated using the principle of geographical weighting (Fotheringham et al. 2002). Objects are geographically weighted during the estimation of each set of local parameter estimates so that more distant objects receive a lower weighting than nearby objects. This results in a geographically local model. Geographically weighted versions of other statistical methods have also been developed on the same principle and these include: GWDA (Brunsdon et al. 2007), Spatio-Temporal Geographically Weighted Regression (STGWR) (Crespo 2009; Huang et al. 2010) and Geographically Weighted Principal Components Analysis (GWPCA) (Harris et al. 2011). Geographically weighted summary statistics have also been developed for both binary (Brunsdon et al. 2002a) and continuous (Brunsdon et al. 2002b) data.

1.2 Geographically weighted discriminant analysis

GWDA is a method for prediction and analysis of categorical spatial data (Brunsdon et al. 2007). The term "discriminant analysis" encompasses a number of related statistical methods that are used to study the typical characteristics of objects belonging to distinct categories (Manly 2005). These characteristics are represented by a set of continuous predictor variables that are used to discriminate between the categories. Each discriminant analysis method makes specific assumptions about the form of the functional relationship between the predictor variables and the categories. For example, Linear Discriminant Analysis (LDA) assumes that the functional relationship is linear whereas Quadratic Discriminant Analysis (QDA) assumes that it is quadratic. These assumptions allow the relationship between the predictor variables and the categories to be modelled mathematically for specific data sets. Discriminant analysis models are typically used, either to predict category membership (classification) or to determine the relative importance of each predictor variable in contributing to category membership (analysis). In this sense, discriminant analysis is similar to regression except that the dependent variable is categorical rather than continuous.

GWDA is a geographically local form of discriminant analysis designed to model a spatially varying or non-stationary relationship between the categories and the predictor variables (Brunsdon et al. 2007). Since spatial non-stationarity does not occur in the physical sciences, this restricts the application of GWDA to social science data.

GWDA models spatial non-stationarity by allowing the parameters of the mathematical function that models the relationship between the predictor variables and the categories to vary spatially (Brunsdon et al. 2007). Note that it is only the parameters of the mathematical function that vary spatially — the functional form (linear or quadratic for example) remains fixed. This results in a geographically local model. Rather than allowing a single function to define the relationship for all geographical locations, each location is now characterised by a unique model. This is accomplished using the principle of geographical weighting (Fotheringham et al. 2002). Objects are geographically weighted during calibration of the GWDA model at each location so that more distant objects receive a lower weighting than nearby objects. In contrast, other discriminant analysis methods such as LDA model stationary relationships and assume that a single model applies regardless of location. The previous analogy between discriminant analysis and regression can also be extended to their geographically weighed equivalents, GWDA and GWR (Fotheringham et al. 2002).

In this thesis, GWDA is developed as a geographically local form of LDA which assumes that the relationship between the predictor variables and the categories is linear. If the relationship between the categories and the predictor variables is really non-stationary, the GWDA model ought to be superior to the LDA model which should result in a higher quality classification. For the purposes of this research, classification quality is assessed in terms of classification uncertainty and classification accuracy.

If GWDA really does improve the quality of the classification, the next question is why? Which variables are most important in determining the assigned categories? Are some variables important in some areas and less important in others? This thesis attempts to provide answers to these questions which will give some insight into the nature of the non-stationary relationship between the categories and the predictor variables.

1.3 Research aims

This thesis aims to develop a methodology to investigate the performance of GWDA using a combination of visual and computational data exploration methods from geovisual analytics. More specifically, the research goals are:

1. The development of a geovisual analytics methodology to assess the GWDA classification quality using the LDA classification quality as a benchmark. For the purposes of this research, classification quality is defined in terms

of classification accuracy and classification uncertainty.

2. The development of a geovisual analytics methodology to explore the extent and nature of spatial non-stationarity in the classification functions. Why does GWDA improve the quality of the classification? Which variables are most important in determining the assigned categories and do these relationships vary spatially?

1.4 Methodology

Addressing these research aims requires analysis of the GWDA and LDA outputs which are characterised by their complex multivariate nature. This research proposes to use geovisual analytic techniques which combine interactive visual tools and computational methods (Andrienko et al. 2007; Kraak 2008) to interpret the GWDA and LDA outputs. Research to date using geovisual analytic techniques to interpret the output of other geographically weighted methods indicates the value of this approach (Dykes and Brunsdon 2007; Demšar et al. 2008a,b; Demšar and Harris 2011). In particular, we develop new visualisation methods (posterior probability treemap, GWDA classification function parameter treemap) and adapt existing methods (legend, thematic map, Parallel Coordinates Plot (PCP), fluctuation diagram) to assist exploration of the GWDA and LDA outputs.

1.5 Case study

Assessment of the effectiveness of the suggested geovisual analytics methodology is provided by a case study. This case study uses GWDA to analyse the relationship between county level voting patterns in the 2004 US presidential election and five socio-economic indicators.

1.6 Thesis structure

This thesis is structured as follows: Chapter 2 introduces GWDA as a non-stationary form of LDA. Chapter 3 explains why techniques from geovisual analytics are useful to interpret the GWDA and LDA output. Chapter 4 describes the new geovisual analytic tools developed to investigate the quality of the GWDA classification and the nature of the spatial non-stationarity. Chapter 5 presents a case study that uses these visualisation tools in an analysis of the relationship between county level voting patterns in the 2004 US presidential election and five socio-economic indicators. Chapter 6 concludes with a discussion the methodology and further suggestions for improvement.

2 Geographically weighted discriminant analysis

2.1 Introduction

This chapter explains the motivation and theoretical foundation for GWDA. First, a general overview of geographically weighted methods is provided since GWDA is a part of the family of geographically weighted methods. Next, the conceptual and mathematical basis for discriminant analysis is explained. Then, the equations for LDA are derived as a prerequisite to the GWDA equations. Although the original GWDA paper by Brunsdon et al. (2007) develops GWDA as an extension of LDA, the authors make it clear that the equations for other discriminant analysis methods, such as Quadratic Discriminant Analysis (QDA) could also be used as a basis. This thesis uses the LDA equations as a basis for GWDA because it is the simplest discriminant analysis method and because the goal of our research is not the development of GWDA per se, but rather the application of techniques from geovisual analytics to investigate the performance of the method. The output from LDA and GWDA is also described in detail since it is used to assess the performance of GWDA. This is required for an understanding of the methodology developed to assess the performance of GWDA which is described in chapter 4.

2.2 Overview of geographically weighted methods

Statistical methods such as linear regression, summary statistics, Principal Components Analysis (PCA) and discriminant analysis all assume that modelled relationships are constant over space. However, in the social sciences, “some relationships are intrinsically different across space” (Fotheringham et al. 2002) and this is referred to as spatial non-stationarity or spatial heterogeneity (Crespo 2009). If a relationship is non-stationary, then a single regression function, summary statistic or PCA transform, calibrated using the entire data set represents a "global average" that will be a poor model of the actual relationship. Geographically weighted methods are one way to model spatially varying relationships. They include: GWR (Fotheringham et al. 2002), geographically weighted summary statistics for binary (Brunsdon et al. 2002a) and continuous (Brunsdon et al. 2002b) data, STGWR (Crespo 2009; Huang et al. 2010), GWPCA (Harris et al. 2011) and GWDA (Brunsdon et al. 2007).

These geographically weighted methods are spatial extensions of the existing aspatial statistical methods: linear regression, summary statistics, PCA and discriminant analysis. Geographically weighted methods take a "geographically local" approach to model spatial non-stationarity. They are "geographically local" in the sense that each geographical location is characterised by a distinct model which is calibrated using the surrounding objects rather than the entire data set. This is accomplished by weighting objects geographically during calibration of the model at each location so that more distant objects have less influence than nearby objects. This is the principal of geographical weighting and underlies all the above geographically weighted methods.

2.3 Discriminant analysis

2.3.1 Overview

The term "discriminant analysis" covers a number of related statistical methods that are used to study the typical characteristics of objects belonging to a set of distinct categories (Manly 2005). The characteristics of objects are represented by a set of continuous predictor variables and these are used to discriminate between the categories. The idea is that objects with similar characteristics should belong to the same category and the discriminant analysis decision rule provides a procedure to assign objects to categories on this basis. This procedure requires a knowledge of the Probability Density Function (PDF) for each category and these are estimated from the data. The PDFs give the conditional probabilities of category membership for each object in the data set. The discriminant analysis decision rule can then be used to assign each object to the category it most closely resembles. Discriminant analysis is a general term for a family of classification methods including: LDA, QDA, Kernel Discriminant Analysis (KDA) and others. The approach is the same in all cases, but the PDFs are estimated differently.

Discriminant analysis is typically used either for supervised classification or to identify the relative importance of the predictor variables in determining the categories (Klecka 1980; McLachlan 2004). Supervised classification requires a training data set. This is comprised of a set of pre-classified objects which can be used to estimate the PDF for each category. The discriminant analysis decision rule can then be used classify new objects. This approach is commonly used for classification of remotely sensed digital imagery. For example, Wilmut et al. (2009) use discriminant analysis to classify a sonar bathymetric data set. In the social sciences, discriminant analysis is typically used to quantify the importance of particular variables in predicting category membership. This approach requires all objects in the data set to have pre-assigned categories. For example, Fotheringham and Reeds (1979) use discriminant analysis to quantify the influence of various socio-economic factors in explaining the choice of alternative crops to tobacco by farmers in Southern Ontario.

2.3.2 Discriminant analysis decision rule

The discriminant analysis decision rule is derived as follows. Suppose there are measurements for m variables on a set of n objects where each object belongs to one of k distinct categories. Each object can be represented by a $m \times 1$ dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_m)$ where the components of $\boldsymbol{x}$ represent the measurements on the m variables.

The goal of discriminant analysis is to assign $\boldsymbol{x}$ to the j^{th} category where $j \in \{1, \ldots, k\}$ so that the probability that $\boldsymbol{x}$ belongs to the j^{th} category, $P(\boldsymbol{x} \cap j)$

is maximized. Bayes' Theorem

$$P(\boldsymbol{x} \cap j) = P(j|\boldsymbol{x})P(\boldsymbol{x}) = P(\boldsymbol{x}|j)P(j) \tag{2.1}$$

is used to compute $P(\boldsymbol{x} \cap j)$ for each $j \in \{1, \ldots, k\}$. $P(j)$, the probability that a randomly selected object belongs to the j^{th} category is known as the prior probability for membership of the j^{th} category and will henceforth be referred to as p_j. Unequal p_js can be used to introduce a natural bias into the decision rule. If there is no natural bias the p_js are assumed to be equal. $P(\boldsymbol{x}|j)$ is equivalent to the PDF, $f_j(\boldsymbol{x})$ for the j^{th} category.

Substituting for $P(\boldsymbol{x}|j)$ and $P(j)$ in equation 2.1 gives the discriminant analysis decision rule. This rule assigns $\boldsymbol{x}$ to the j^{th} category where

$$\{C_j(\boldsymbol{x}) = p_j f_j(\boldsymbol{x}) : j \in \{1, \ldots, k\}\} \tag{2.2}$$

is maximized.

2.4 Linear discriminant analysis

2.4.1 Linear discriminant analysis assumptions

The LDA equations are derived from the discriminant analysis decision rule given by equation 2.2 under the following assumptions:

1. The PDF for each category, $f_j(\boldsymbol{x})$ is multivariate normal.

2. All categories have the same covariance matrix, Σ.

This results in the LDA decision rule which is linear in $\boldsymbol{x}$.

2.4.2 Linear discriminant analysis equations

The first assumption of LDA (see section 2.4.1) states that the PDF for each category is multivariate normal. The multivariate normal PDF $f_j(\boldsymbol{x})$, for the j^{th} category is given by

$$f_j(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}|\Sigma_j|^{1/2}} \ \mathrm{e}^{\left[-1/2(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right]} \tag{2.3}$$

where the $\boldsymbol{\mu}_j = (\mu_{j,1}, \ldots, \mu_{j,m})$ is the mean vector for the j^{th} category, Σ_j is the covariance matrix for the j^{th} category and e^x is the exponential function (Sharma 1996).

Substituting the multivariate normal PDF (equation 2.3) into the discriminant analysis decision rule (equation 2.2) and taking the natural logarithm[1] of the result gives a decision rule that assigns $\boldsymbol{x}$ to the j^{th} category where

$$\left\{-\frac{m}{2}\ln|\Sigma_j| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) + \ln(p_j)\right\} \quad \forall \quad j \in \{1, \ldots, k\} \tag{2.4}$$

is maximized.

The second assumption of LDA (see section 2.4.1) states that each category has the same covariance matrix, Σ. Substituting Σ for Σ_j into equation 2.4 and simplifying[2], results in the LDA decision rule

$$\left\{C_j(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) + \ln(p_j)\right\} \quad \forall \quad j \in \{1, \ldots, k\}. \tag{2.5}$$

Brunsdon et al. (2007) suggest estimating Σ as a weighted average of the individual category covariance matrices Σ_j so that

$$\Sigma = \frac{n_1\Sigma_1 + n_2\Sigma_2 + \ldots + n_k\Sigma_k}{n_1 + n_2 + \ldots + n_k} \tag{2.6}$$

[1] Taking the natural logarithm works because it is a monotonically increasing function and thus preserves the order in the domain.

[2] Note that the expression $-\frac{m}{2}\ln|\Sigma|$ is common to all categories and can therefore be omitted from the LDA decision rule.

and where n_j is the number of objects in the j^{th} category.

The LDA decision rule in equation 2.5 can also be rewritten as a linear combination of the components of $\boldsymbol{x}$

$$\left\{C_j(\boldsymbol{x}) = \boldsymbol{x}'\Sigma^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j'\Sigma^{-1}\boldsymbol{\mu}_j + \ln(p_j)\right\} \quad \forall \quad j \in \{1, \ldots, k\}. \tag{2.7}$$

These are the LDA classification functions. The parameters of these functions describe the relationship between the categories and the attributes. $\boldsymbol{x}$ is classified by assigning it to the category $j \in \{1, \ldots, k\}$ with the greatest classification function value, $C_j(\boldsymbol{x})$.

Note that classification by maximizing the j^{th} classification function in equation 2.7 is equivalent to assigning $\boldsymbol{x}$ to the category that minimizes the Mahalanobis distance squared

$$\left\{D_j(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_j)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)\right\} \quad \forall \quad j \in \{1, \ldots, k\} \tag{2.8}$$

between $\boldsymbol{x}$ and $\boldsymbol{\mu}_j$ if the p_js are equal. The Mahalanobis distance squared is a multivariate distance measurement (Manly 2005).

Note also from equation 2.1 (Bayes' Theorem) that an equivalent result is obtained by maximizing $P(j|\boldsymbol{x})$ for $j \in \{1, \ldots, k\}$. $P(j|\boldsymbol{x})$ is the posterior probability that $\boldsymbol{x}$ belongs to the j^{th} category. This is the probability that $\boldsymbol{x}$ belongs to the j^{th} category given that $\boldsymbol{x}$ must belong to one of the categories $j \in \{1, \ldots, k\}$. Therefore, the sum of the k posterior probabilities for each $\boldsymbol{x}$ is equal to one. $P(\boldsymbol{x})$ is given by $\sum_{i=1}^{k} p_i f_i(\boldsymbol{x})$. The k posterior probabilities for $\boldsymbol{x}$ are given by

$$\left\{P(j|\boldsymbol{x}) = \frac{p_j f_j(\boldsymbol{x})}{\sum_{i=1}^{k} p_i f_i(\boldsymbol{x})}\right\} \quad \forall \quad j \in \{1, \ldots, k\} \tag{2.9}$$

and represent the confidence or uncertainty of the classification. Although the classification function values are related to the posterior probabilities by a scaling of $P(\boldsymbol{x})$, they are only suitable for classification and cannot be compared to determine the confidence of the classification since the range of possible classification function values is not fixed.

The next section explains how the LDA output is used to determine the classification accuracy, classification uncertainty and the relative importance of each variable in the classification in the context of the research aims given in section 1.3.

2.4.3 Linear discriminant analysis output

The output from LDA comprises:

1. The n assigned categories.
2. The posterior probabilities $P(j|\boldsymbol{x})$ for category membership where $j \in \{1, \ldots, k\}$. There are k posterior probabilities for each object $\boldsymbol{x}$.
3. The Mahalanobis distance squared $D_j(\boldsymbol{x})$ from each observation to each category mean where $j \in \{1, \ldots, k\}$. There are k distances for each object $\boldsymbol{x}$.
4. The classification functions parameters. There are $m + 1$ of these for each of the k classification functions.

The LDA classification accuracy can be determined by comparing the n assigned categories with the n actual categories. The uncertainty of the LDA classification can be examined by analysis of the posterior probabilities. Finally, the LDA classification function parameters represent the relative importance of each predictor variable in the classification. However, since there are k classification function parameters for each predictor variable, interpretation is difficult. Additionally, if the predictor variables are scaled differently, this will also affect interpretation of the parameters. In this thesis, the predictor variables are standardised prior to the classification to compensate for this (see sections 4.7 and 5.4.4).

2.5 Geographically weighted discriminant analysis

2.5.1 Motivation

GWDA (Brunsdon et al. 2007) is an extension of LDA that enables the prediction and analysis of categorical spatial data where the relationship between the predictor variables and the categories varies spatially. This is also referred to as spatial non-stationarity or spatial heterogeneity. If the relationship between the predictor variables and the categories is non-stationary, GWDA would be expected to more accurately model the relationship between the predictor variables thus providing a higher quality classification than LDA.

2.5.2 Geographical weighting and spatial non-stationarity

GWDA models spatial non-stationarity by allowing the parameters of the LDA classification functions (see equation 2.7) to vary spatially (Brunsdon et al. 2007). This is accomplished by allowing some or all of $\boldsymbol{\mu}_j$, Σ and p_j to be functions of geographical space $\boldsymbol{u} = (a, b)$ where a is the coordinate along the horizontal axis (the easting) and b is the coordinate along the vertical axis (the northing). Local estimates at $\boldsymbol{u}$ for each of these quantities are computed from objects in the vicinity of $\boldsymbol{u}$. Objects are geographically weighted in the calculation of the local estimates according to their geographic distance from $\boldsymbol{u}$ so that distant objects are less influential than nearby objects. The distance from $\boldsymbol{u}$ where the weight falls to zero is referred to as the *bandwidth*, h. h may be constant (*fixed*) or spatially varying (*adaptive*). In the latter case, h is a function of $\boldsymbol{u}$ and $h(\boldsymbol{u})$ expands and contracts depending on the local spatial density of the objects. An adaptive bandwidth allows $h(\boldsymbol{u})$ to contract where objects are geographically dense and expand where there are fewer objects. $h(\boldsymbol{u})$ is estimated by specifying that each local estimate must contain a minimum of N nearest neighbours in each category.

The geographical weighting of local estimates to model spatial non-stationarity in GWDA is conceptually identical to the other geographically weighted methods discussed in section 2.2.

Brunsdon et al. (2007) suggest a bisquare kernel weighting function to assign geographical weights, w_i to the objects surrounding the object currently being classified, $\boldsymbol{x}$. These are given by

$$w_i(\boldsymbol{u}) = \begin{cases} (1 - (d_i/h)^2) & \text{if} \quad d_i < h \\ 0 & \text{if} \quad d_i \geq h \end{cases} \tag{2.10}$$

where d_i is the geographic distance between $\boldsymbol{u}$ and the i^{th} object in the data set.

Using the bisquare kernel weighting function given by equation 2.10, the geographically weighted category means $\boldsymbol{\mu}_j(\boldsymbol{u})$, pooled category covariance matrix $\Sigma(\boldsymbol{u})$ and prior probabilities $p_j(\boldsymbol{u})$ can be calculated at each geographical location $\boldsymbol{u}$. Inserting these in the LDA classification functions given by equation 2.7 results in a set of classification functions specific to each location $\boldsymbol{u}$. These are the GWDA classification functions.

2.5.3 Geographically weighted discriminant analysis equations

The geographically weighted mean for the j^{th} category, $\boldsymbol{\mu}_j(\boldsymbol{u})$ is given by

$$\boldsymbol{\mu}_j(\boldsymbol{u}) = \frac{\displaystyle\sum_{\boldsymbol{x}_i \in j} w_i(\boldsymbol{u})\boldsymbol{x}_i}{\displaystyle\sum_{\boldsymbol{x}_i \in j} w_i(\boldsymbol{u})} \tag{2.11}$$

where $\boldsymbol{x}_i$ is the i^{th} object belonging to the j^{th} category and $w_i(\boldsymbol{u})$ is the geographical weight applied to the i^{th} object. There are k geographically weighted means at each geographical location $\boldsymbol{u}$.

The geographically weighted covariance matrix for the j^{th} category, $\Sigma_j(\boldsymbol{u})$ is given by

$$\Sigma_j(\boldsymbol{u}) = \frac{\sum\limits_{\boldsymbol{x}_i \in j} w_i(\boldsymbol{u}) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j(\boldsymbol{u})\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j(\boldsymbol{u})\right)'}{\sum\limits_{\boldsymbol{x}_i \in j} w_i(\boldsymbol{u})} \tag{2.12}$$

where $\boldsymbol{x}_i$ is the i^{th} object belonging to the j^{th} category, $w_i(\boldsymbol{u})$ is the geographically varying weight applied to the i^{th} object and $\boldsymbol{\mu}_j(\boldsymbol{u})$ is the geographically weighted mean for the j^{th} category. There are k geographically weighted covariance matrices at each geographical location $\boldsymbol{u}$.

The geographically weighted pooled covariance matrix $\Sigma(\boldsymbol{u})$ is estimated by computing a weighted average of the k geographically weighted covariance matrices in equation 2.12 according to equation 2.6.

The geographically weighted prior probabilities for the j^{th} class, $p_j(\boldsymbol{u})$ are given by

$$p_j(\boldsymbol{u}) = \frac{\sum\limits_{\boldsymbol{x}_i \in j} w_i(\boldsymbol{u})}{\sum\limits_{j} \sum\limits_{\boldsymbol{x}_i \in j} w_i(\boldsymbol{u})} \tag{2.13}$$

where $w_i(\boldsymbol{u})$ is the geographically varying weight applied to the i^{th} object belonging to the j^{th} category. There are k geographically weighted prior probabilities at each geographical location $\boldsymbol{u}$.

Cross-validation is used to identify the optimum number of nearest neighbours, N_o. This involves calibrating the k GWDA classification functions at each location $\boldsymbol{u}$ using the set of all objects in the data set, but excluding the object at $\boldsymbol{u}$. A range of different N is specified and N_o is that which maximizes either the proportion of correct classifications, or the sum of the logs of the posterior probabilities.

The next section describes how the GWDA output differs from the LDA output described in section 2.4.3. It also explains how the GWDA output is used to determine the classification accuracy, classification uncertainty and the relative importance of each variable in the classification in the context of the research aims given in section 1.3.

2.5.4 Geographically weighted discriminant analysis output

The output from GWDA comprises:

1. The n assigned categories.

2. The posterior probabilities $P(j|\boldsymbol{x})$ for category membership where $j \in \{1, \ldots, k\}$.. There are k of these for each $\boldsymbol{x}$.

3. The Mahalanobis distance squared $D_j(\boldsymbol{x})$ from each observation to each category mean where $j \in \{1, \ldots, k\}$. There are k of these for each $\boldsymbol{x}$.

4. The spatially varying classification functions parameters. There are $m + 1$ of these for each of the k classification functions which gives a total of $(m + 1) \times k$ parameter surfaces.

5. The spatially varying bandwidth function $h(\boldsymbol{u})$.

6. The spatially varying category means. There are k of these surfaces.

7. The spatially varying covariance matrix. There are $\frac{m(m+1)}{2}$ of these surfaces.

The first three items on this list are identical to the LDA output described in section 2.4.3 in terms of complexity, although they will differ numerically if non-stationarity is present. The final four items are unique to GWDA and add significantly to the complexity of the output.

One way to make sense of complex geospatial data sets is through visualisation (Keim et al. 2003) and that is the approach followed in this thesis. To date, the only attempt at visualising the GWDA output is contained in the paper by Brunsdon et al. (2007). They note that simply following the GWR approach (Fotheringham et al. 2002) and mapping each of the classification function parameters as separate univariate choropleth maps is likely to result in information overload. Instead, they suggest mapping the spatial variation in posterior probabilities for a fixed set of predictor variables. A distinctive hue is assigned to each category and the map is coloured, either by the hue of the assigned category or by a mix of hues in proportion to the posterior probabilities. The difficulty with the latter approach is perceptual, since it assumes that people are able to decompose a colour formed by mixture of hues into the original hues and in the correct proportions. We illustrate this with the following example. Figure 2.1 shows three hues used to represent three categories chosen optimally from the CIELUV colour space. These hues were chosen to be as far apart and thus as distinguishable as possible. The CIELUV colour model is a perceptually uniform colour model developed by the Commission Internationale de l'Eclairage (CIE) and uses three coordinates: L,U and V to represent colours (Slocum et al. 2009). Equal distances in (L,U,V) space correspond to how humans perceive the change in colour (Slocum et al. 2009). The three hues in figure 2.1 are defined for a fixed value of L (luminance) since Brunsdon et al. (2007) note that zones on a map with greater luminance values dominate. Thus the three hues in figure 2.1 are chosen from the (U,V) plane only. Since people with normal colour vision are

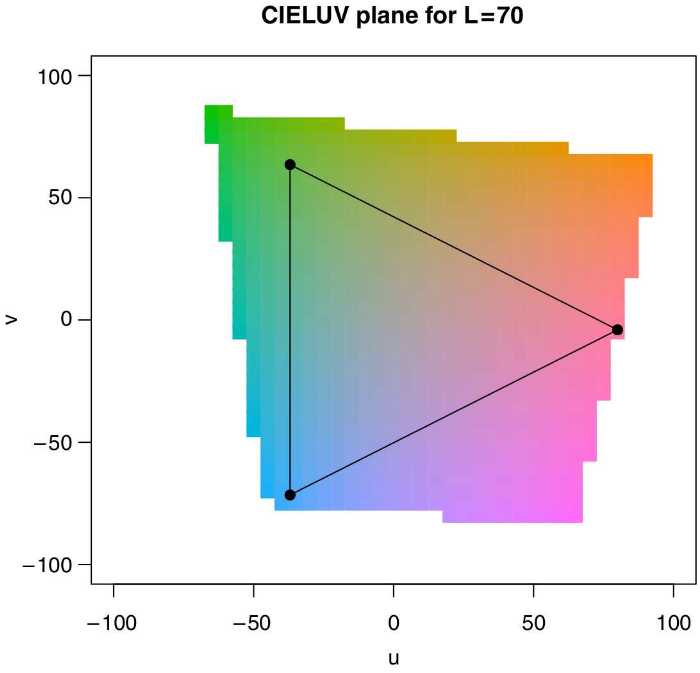


Figure 2.1: Using the CIELUV colour space to visualise a three-way probability vector. *Source*: Reprinted with permission of Prof. Chris Brunsdon from Geographically Weighted Discriminant Analysis. *Geographical Analysis* 2007;39(4):376-396.

naturally sensitive to only three distinct colours (Slocum et al. 2009), colouring by a mix of hues in proportion to the posterior probabilities is likely to be difficult for more than three categories. A different, more general approach is therefore required.

In a case study to demonstrate the effectiveness of GWDA, Brunsdon et al. (2007) present another visualisation approach and colour maps by assigned category. They fix the predictor variables at three levels: one standard deviation below the mean, the mean and one standard deviation above the mean and use small map multiples to show spatial variation in assigned categories for all combinations. This spatial variation is used as evidence for non-stationarity. A disadvantage of the small multiple approach is that comparison of two geographical areas across a set of attributes is difficult (Slocum et al. 2009). In the case of GWDA, 3^m maps are required. This works for Brunsdon et al. since they use two predictor variables which requires nine maps. However, the same approach in our case-study would require $3^5 = 243$ maps. This is not practical and again, a different approach is required.

To solve these two problems, this thesis suggests a new approach that uses techniques from the field of geovisual analytics.

3 Geovisual analytics

3.1 Introduction

This thesis proposes to use techniques from geovisual analytics to investigate the performance of GWDA. This chapter presents a review of the geovisual analytics literature in the context of this research. The concept of and motivation for geovisual analytics is explained and the historical background discussed. Example applications of geovisual analytic methods from the literature are provided, with an emphasis on applications that combine geovisual analytics and geographically weighted methods. Finally, a justification for the use of geovisual techniques to address the research aims (see section 1.3) is provided.

Geovisual analytics is the sub-discipline of visual analytics that deals with spatial data (Andrienko et al. 2007). Visual analytics has been defined as combining 'automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets' (Keim et al. 2010).

Geovisual analytics involves the development of tools and techniques that combine computational methods with interactive visual representations to *make sense of* large complex multivariate spatio-temporal data sets. In this context, *make sense of* refers to the transformation of data into information and ultimately into new knowledge about the processes captured by the data. The transformation of data into information involves the detection of unexpected and hidden patterns in the data. These patterns are interpreted by analysts in the context of background knowledge and these interpretations are used to form new hypotheses. Testing these hypotheses results in new knowledge about the processes captured by the data. The ultimate aim of geovisual analytic methods is to help users *gain insight* into the data (Kang et al. 2011).

A very simple example of this is the use of the correlation coefficient with a two-dimensional scatter plot to identify the strength of the linear relationship between two variables. The correlation coefficient is the result of a computational method and quantifies the extent of a linear relationship between two variables. However, if the relationship between the two variables is really non-linear the information provided by the correlation coefficient could be misleading. Instead, a display of both variables on a two-dimensional scatter plot to show the relationship between them graphically can be used to hypothesize about the actual mathematical relationship (see figure 3.1).

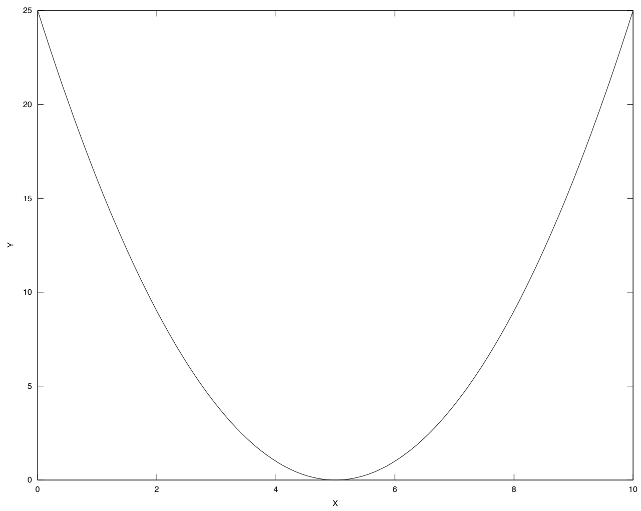


Figure 3.1: Plot of $f(x) = x^2$. The relationship between the variables x and y is mathematically precise (parabolic) but the correlation coefficient is 0.

The motivation for geovisual analytics lies in the *information overload* problem posed by the increasingly large and complex multivariate spatio-temporal data sets generated by researchers and industry. Existing visualisation techniques are insufficient due to the magnitude and complexity of these data sets as well as the ill-defined nature of exploratory data analysis (Keim et al. 2004; Andrienko et al. 2007). Combining automatic methods with interactive visual representations of the data helps 'foster new insights and encourages the formation and validation of new hypotheses' (Keim et al. 2004).

3.2 Historical background

Although the terms ‘visual analytics’ and ‘geovisual analytics’ are relatively new — the term ‘visual analytics’ has only been in use since 2005 (Andrienko et al. 2010), the combination of computational methods with visualization tools is somewhat older. Shneiderman (2002) proposed combining statistical/data mining algorithms with visual tools for exploratory data analysis. The goal of this process is hypothesis generation and is very similar to what is today described as visual analytics. Keim et al. (2003, 2004) proposed combining data mining algorithms with interactive visual tools to search for patterns in large geo-spatial data sets. Keim et al. (2004) believed that fully automated data mining algorithms often produced unsatisfactory results and that a synthesis of automatic and visual methods would not only give better results, but lead to a higher degree of user confidence in the results. This approach is analogous to geovisual analytics.

Geovisualization, an older term again, also deals with similar issues to geovisual analytics. MacEachren and Kraak (2001) identified the integration of visual and computational methods, data representation, interface design and cognitive issues as being crucial themes of the geovisualization research agenda. Dykes et al. (2005) also stress the interdisciplinary aspect of geovisualization which they describe as combining the disciplines of cartography, scientific visualisation, image analysis, information visualisation, exploratory data analysis, and GIScience.

In the United States of America (US), the visual/geovisual analytics research agenda is dominated by national security concerns as a result of the terrorist attacks in New York on 11th, 2001. A significant challenge in preventing future terrorist attacks lies in the analysis of large multi-dimensional multi-source temporal data sets. These data sets might include details such as immigration records, patterns of travel, telephone calls, names, affiliations and locations (for example). The analytic goal is to identify patterns that might suggest terrorist activity. The National Visualization and Analytics Center (NVAC) was established in 2004 by the US Department of Homeland Security with the goal of helping to counter future terrorist attacks (Thomas and Cook 2005). In 2005, NVAC published the book 'Illuminating the Path' (Thomas and Cook 2005) which defined the term 'visual analytics' and provided a five year research agenda. Although the research agenda was driven by issues of national security, the authors did acknowledge that visual analytic techniques would have a significant impact on other research areas (Thomas and Cook 2005).

The European geovisual analytics research agenda is driven by the VisMaster consortium (`http://www.vismaster.eu`) but is not restricted to issues of national security or disaster management. The European research agenda includes any problem with a spatial component. It also stresses the importance of both the temporal and spatial components together (Andrienko et al. 2010). As an application of geovisual analytics, Andrienko et al. (2007) define spatial decision support as "computerized assistance to people in the development, evaluation, and selection of proper policies, plans, projects or interventions where the problems have a geographic or spatial component". Spatial decision support includes more sophisticated multi-criterial decision analysis methods for site selection, forestry, emergency response or hazard avoidance (Andrienko et al. 2007, 2010). The European visual analytics research agenda "Mastering the information age: solving problems with visual analytics" (Keim et al. 2010) was published in 2010.

There are diverse applications of geovisual analytic techniques in the literature including: archaeology (Huisman et al. 2009), thermography (Danese et al. 2009), climatology (Steed et al. 2009), remote sensing (Ahmed et al. 2009), textual analysis (Tomaszewski et al. 2011; Luo et al. 2012) and analysis of geo-tagged social networking data (Jankowski et al. 2010), to name a few examples.

3.3 Geovisual analytics and spatial statistics

Geovisual analytic methods have been successfully used to interpret the output of various spatial statistical methods. Demšar et al. (2008a) use a geovisual analytics environment to identify multivariate spatial and non-spatial relationships and patterns in the output of GWR. Their geovisual analytics environment was constructed using GeoVISTA Studio (Gahegan et al. 2002) and uses a self-organizing map, a geographic map and a PCP to identify clusters of observations in parameter space. Demšar et al. (2008b) also demonstrate the ability of geovisual analytics methods to discover expected patterns in the output of STGWR. Demšar and Harris (2011) use geovisual analytic tools from the GeoViz Toolkit (Hardisty and Robinson 2011) to evaluate and assess the performance of moving window kriging, a non-stationary spatial prediction method. Dykes and Brunsdon (2007) combined geographically weighted local statistics and interactive visualizations (so-called *geowigs*) and used them to explore a multivariate spatial data set.

For these reasons, it is anticipated that the application of geovisual analytic methods to the GWDA output has the potential to improve on the existing techniques described in section 2.5.4.

4 Methodology

4.1 Introduction and software

This section describes the interactive geovisual analytic tools that have been developed to investigate the performance of GWDA (see figure 4.1). These tools are demonstrated in a case study that is described in chapter 5.

Figure 4.1 consists of the following visualisations (top left to bottom right):

- Interactive legends for the actual categories, assigned LDA and GWDA categories, the five predictor variables, the GWDA classification uncertainty and the GWDA classification function parameters.
- A treemap to explore the GWDA posterior probabilities.
- Two thematic maps showing the spatial distribution of categorical or continuous variables. These maps can be used to show the spatial distribution of the actual categories, assigned LDA categories, assigned GWDA categories and the five predictor variables.
- A treemap to explore the significance of the GWDA classification function parameters for each object in the data set.
- A PCP to investigate the relationship between the actual categories and predictor variables.
- Fluctuation diagrams to visualise the LDA and GWDA confusion matrices.
- A slider to detect outliers at different significance levels.

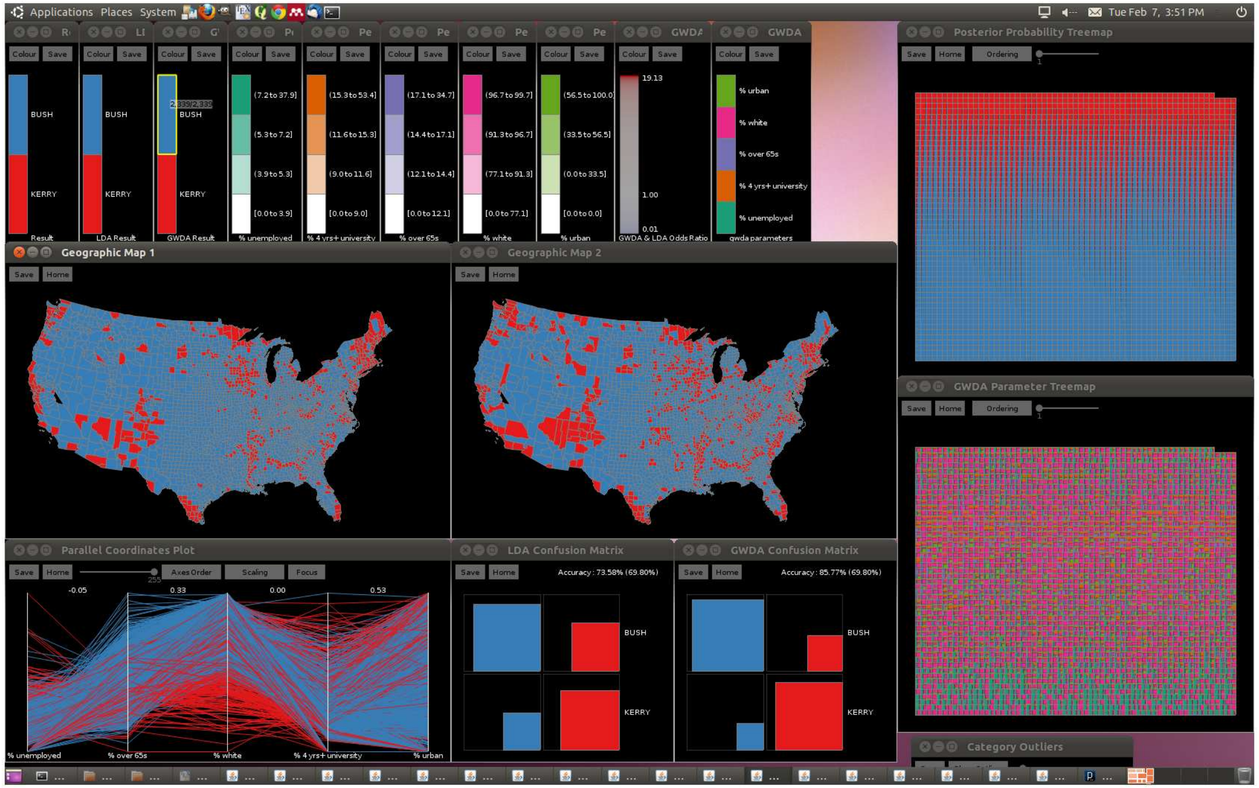
The design of these tools is based on the structure of the existing geovisual analytics environments, *GeoVISTA Studio* (Gahegan et al. 2002) and the *GeoViz Toolkit* (Hardisty and Robinson 2011). The geovisual analytic tools developed for this research use multiple linked views, dynamic interaction with selection and brushing and additional methodologies for statistical analyses of spatial non-stationarity, classification uncertainty and detection of outliers. Details of the

1. Interactive legends for the actual categories, assigned LDA and GWDA categories, the five predictor variables, the GWDA classification uncertainty and the GWDA classification function parameters

3. Thematic maps showing the spatial distribuion of the actual and assigned categories

5. PCP to explore the relationship between the categories and the predictor variables

6. Fluctuation diagrams visualising the LDA and GWDA confusion matrices



2. Interactive treemap to explore the GWDA posterior probabilities

4. Interactive treemap to explore the significance of the GWDA classification function parameters

7. A slider to detect outliers at different significance levels

Figure 4.1: System of interactive, linked geovisual analytic tools for investigating the performance of GWDA. The system consists of the following linked visualisations (top left to bottom right): 1. Interactive legends for the actual categories, assigned LDA and GWDA categories, the five predictor variables, the GWDA classification uncertainty and the GWDA classification function parameters; 2. An interactive treemap to explore the GWDA posterior probabilities; 3. Two thematic maps showing the spatial distribution of the actual and assigned categories; 4. An interactive treemap to explore the significance of the GWDA classification function parameters; 5. A PCP to investigate the relationship between the categories and predictor variables; 6. Fluctuation diagrams visualising the GWDA and LDA confusion matrices; 7. A slider to detect outliers at different significance levels.

software libraries used to construct the visualisations are provided in the appendices.

These geovisual analytic tools implement the seven basic ‘tasks’ for information visualisation as far as possible (Shneiderman 1996):

- **Overview:** The default for each view is to provide an overview of all objects in the data set.
- **Zoom:** Interactive panning and zooming is available on all views.
- **Filter:** Interactive selection using the mouse is available on all views. Traditional boolean selection where objects must be either ‘selected’ or ‘unselected’ is restrictive when exploring the conditional distributions of categorical data. Therefore, a method of ternary selection is proposed instead. This means an object must be in one of three states:
  - ‘Hidden’. These objects do not appear in any view nor are they used in statistical calculations. This feature is activated using a hot-key.
  - ‘Unselected’. These objects are coloured pale grey in all views. They are equivalent to ‘unselected’ objects using boolean selection.
  - ‘Selected’. These objects are in colour and placed above ‘unselected’ objects. They are equivalent to ‘selected’ objects using boolean selection.

  Ternary selection is based on the concept of *Degree of Interest* (Wills 2008). Ternary selection allows the user to focus initially on a particular subset of the data by hiding non-relevant data and create further selections within this subset. This feature is particularly useful when exploring the nature of the differences between the correctly classified and misclassified objects for a specific category.
- **Details-on-demand:** When objects in different views are brushed with the mouse, the objects are highlighted and pop-up labels appear with detailed

information about them.

- **Relate:** All views are linked and communicate using events. The following events have been defined; Colour Events, Brushing Events and Degree of Interest events.
- **History:** A history type mechanism has not been implemented.
- **Extract:** It is possible to save high quality pdf images of each view to capture the visual representation of identified patterns.

In the remainder of this chapter, each of the geovisual analytics tools shown in figure 4.1 is described in detail, together with the tasks for which they were developed in the context of GWDA.

4.2 Interactive legends and thematic maps

The interactive legends shown in figure 4.2, together with the dynamically linked thematic maps in figure 4.4 are used to explore the spatial distribution of the actual categories, the assigned LDA and GWDA categories and the predictor variables.

The legend in figure 4.2(a) uses categorical data, in this case the outcome of the 2004 US presidential election. The five classed legends in figures 4.2(b), 4.2(c), 4.2(d), 4.2(e) and 4.2(f) use continuous data, classified into quartiles. These five variables are the predictor variables for the LDA and GWDA classifiers used in the case study that is described in chapter 5.

The system of linked geovisual analytic tools shown in figure 4.1 uses two thematic maps. One of these is dynamically linked to the categorical legend and the other is linked to the predictor variable legends. Clicking the *Colour* button in the top left-hand corner of each legend updates the linked thematic map with the associated colour scheme in real time, allowing rapid comparison of the spatial distributions of the categories and the predictor variables.

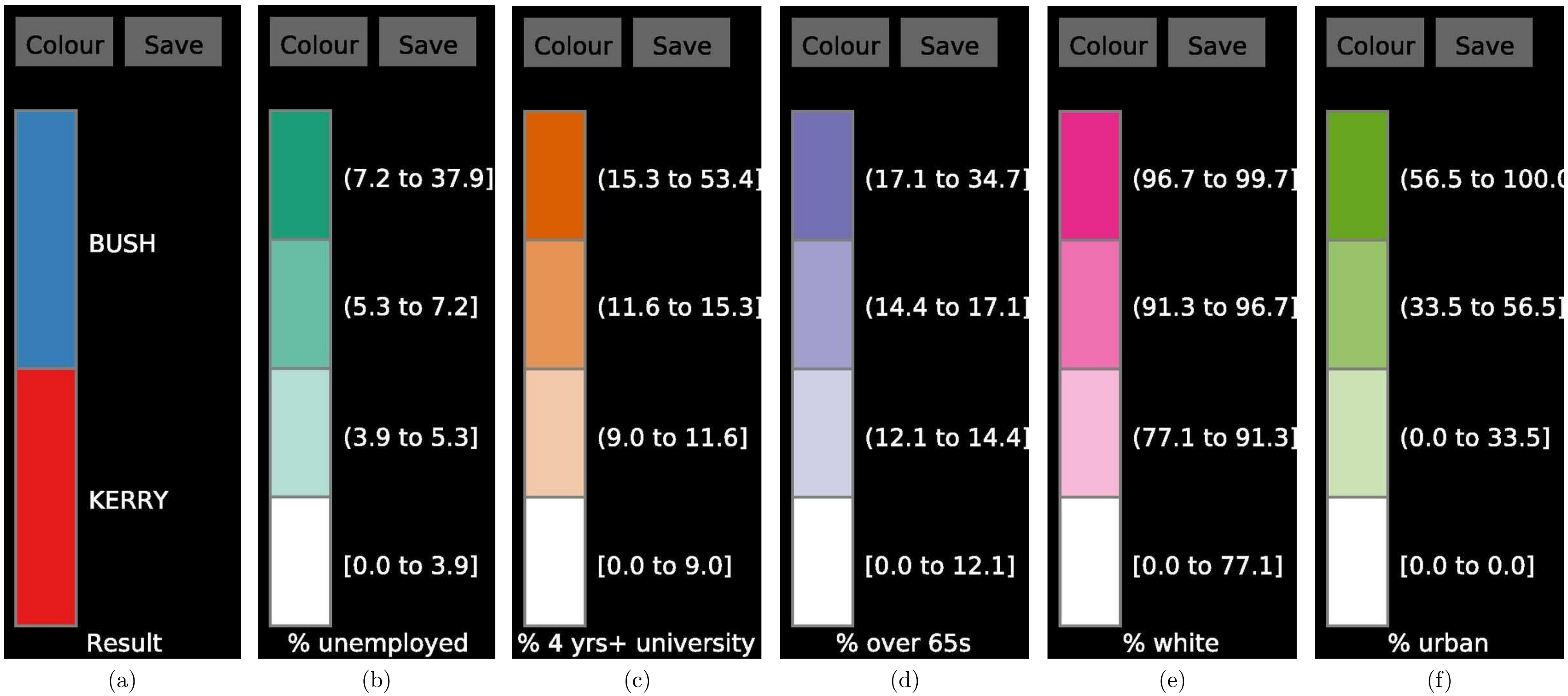


Figure 4.2: Interactive legends for categorical and continuous data. Figure 4.2(a) is a categorical legend showing the colours assigned to the 2004 US presidential election results. Figures 4.2(b), 4.2(c), 4.2(d), 4.2(e) and 4.2(f) are legends for the five predictor variables used to predict the election results in the GWDA and LDA models (*% unemployed*, *% adults over* 25 *with four or more years of college education*, *% persons over* 65, *% white* and *% urban*). Each of these five legends is classified into quartiles, which form the legend categories. Note that the bottom quartile for *% urban* in figure 4.2(f) comprises 826 counties with no urban area.

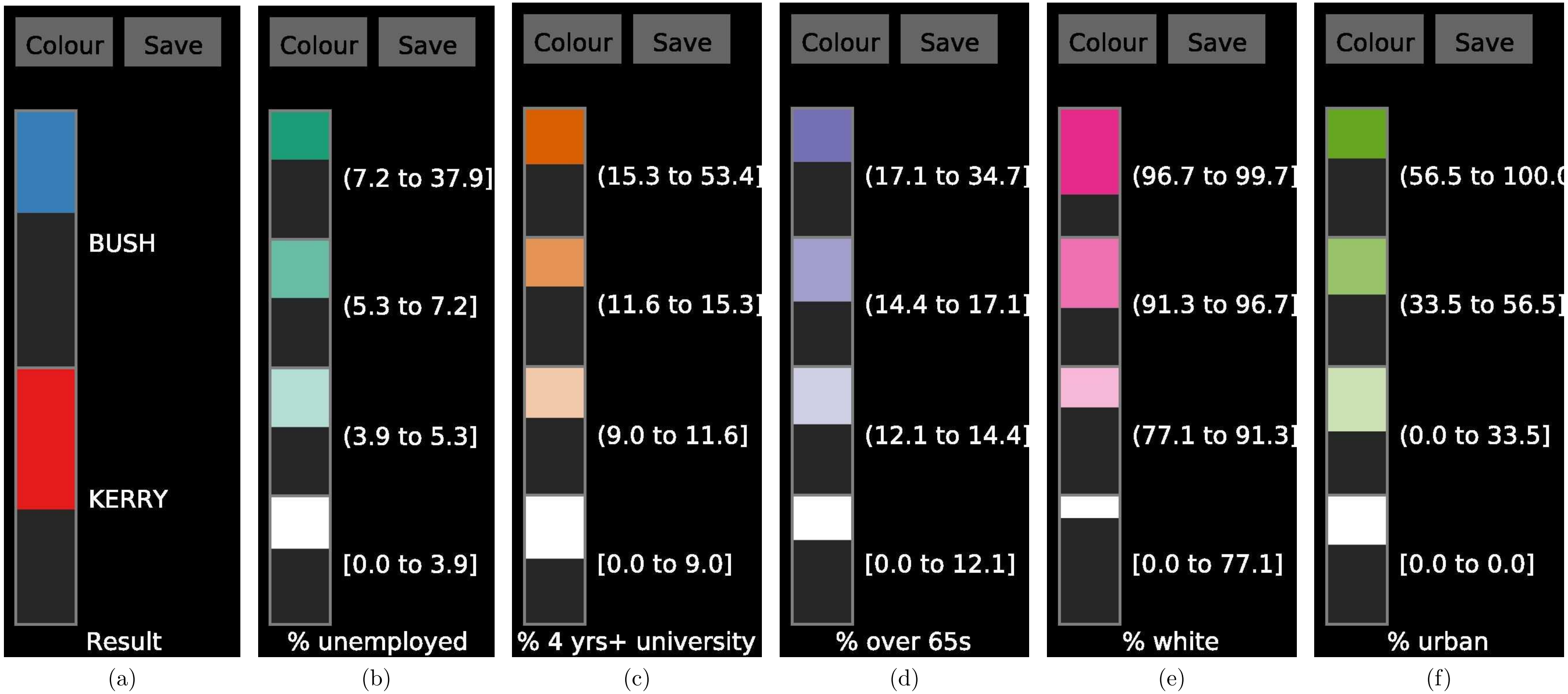
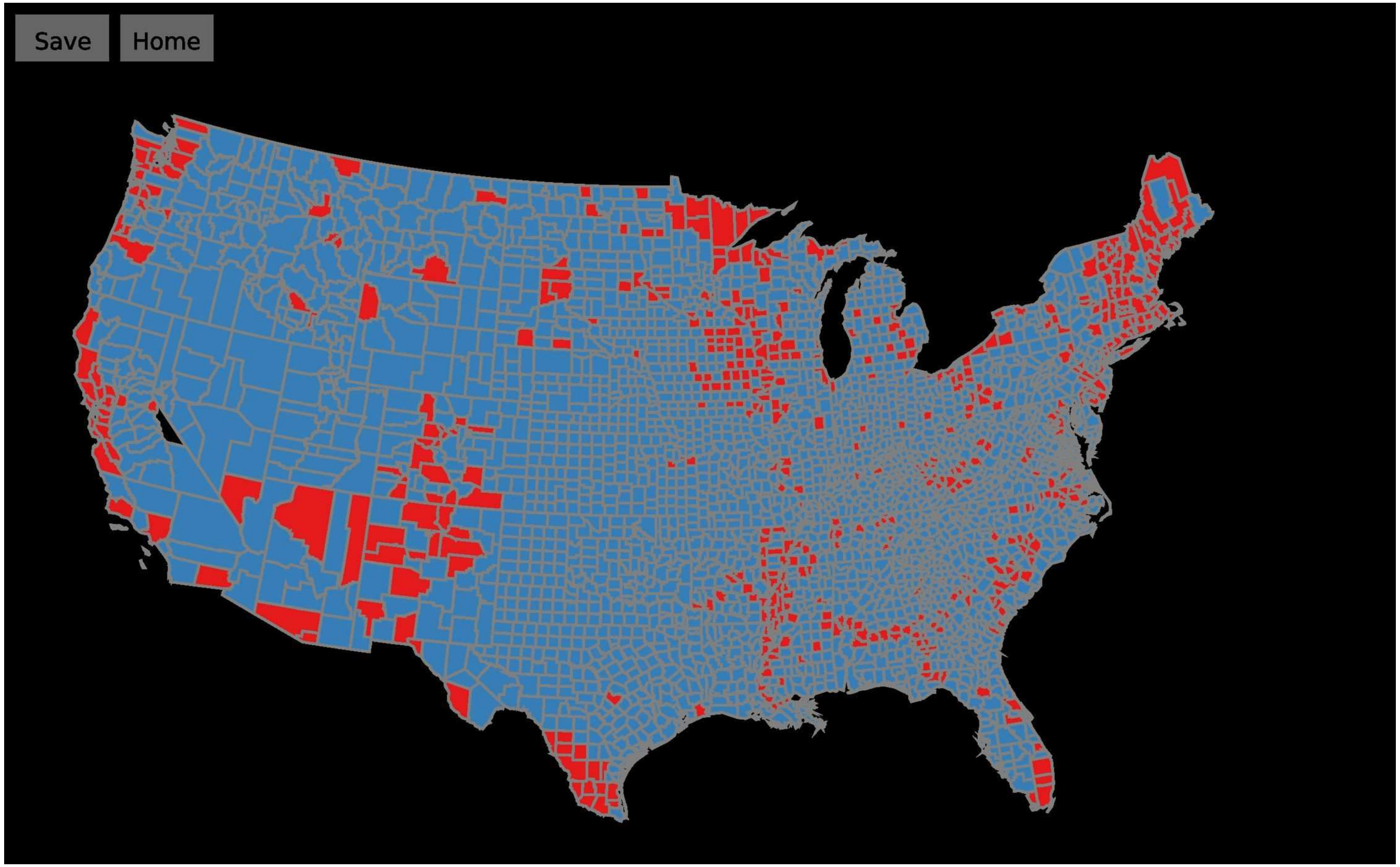
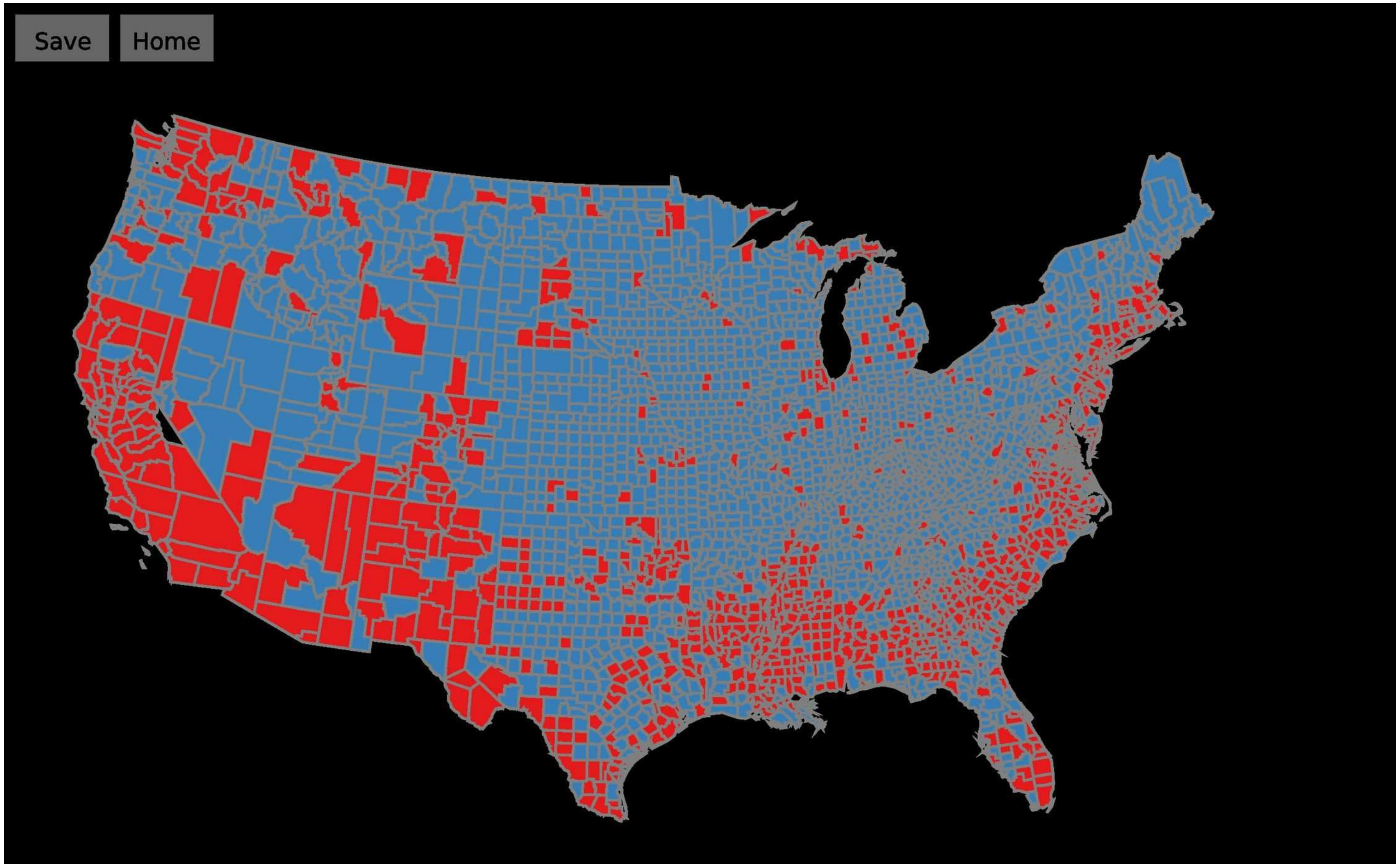


Figure 4.3: Interactive legends for categorical and continuous data showing a subset of the data, selected in a dynamically linked thematic map. These legends are identical to the legends in figure 4.2. The height of each legend class is proportional to the number of objects selected. Unselected objects are shown in grey.



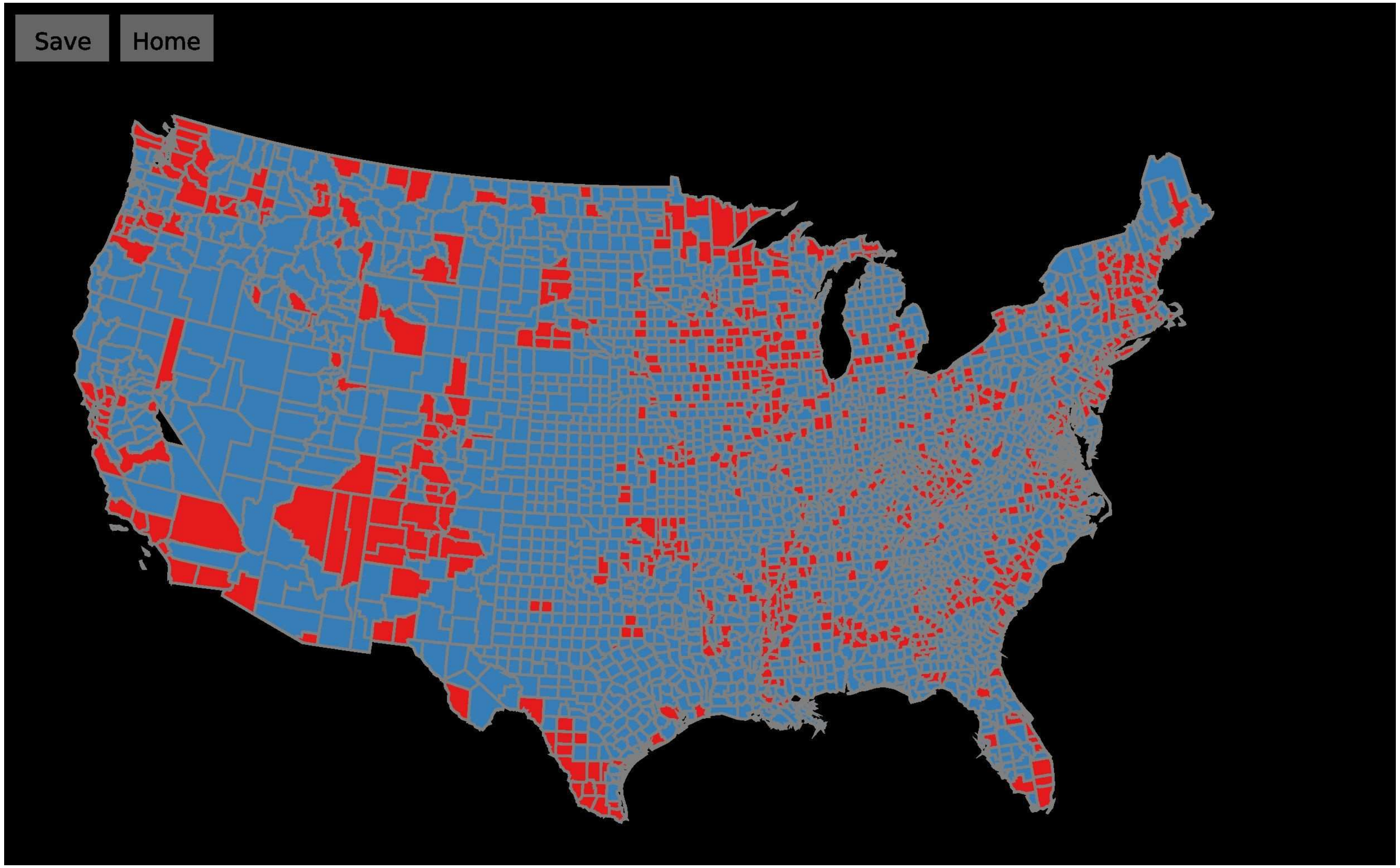
(a) Thematic map showing the spatial distribution of the actual 2004 US presidential election results by county.

Figure 4.4: Thematic maps showing the spatial distribution of the actual categories, the LDA assigned categories and the GWDA assigned categories. The categories are the county level 2004 US presidential election results. The maps are coloured according to the legend in figure 4.2(a) so that counties won by George Bush are coloured blue and counties won by John Kerry are coloured red.



(b) Thematic map showing the spatial distribution of the LDA assigned 2004 US presidential election results by county.

Figure 4.4: Thematic maps showing the spatial distribution of the actual categories, the LDA assigned categories and the GWDA assigned categories. The categories are the county level 2004 US presidential election results. The maps are coloured according to the legend in figure 4.2(a) so that counties won by George Bush are coloured blue and counties won by John Kerry are coloured red.



(c) Thematic map showing the spatial distribution of the GWDA assigned 2004 US presidential election results by county.

Figure 4.4: Thematic maps showing the spatial distribution of the actual categories, the LDA assigned categories and the GWDA assigned categories. The categories are the county level 2004 US presidential election results. The maps are coloured according to the legend in figure 4.2(a) so that counties won by George Bush are coloured blue and counties won by John Kerry are coloured red.

Selection of individual classes within each legend is possible by mouse-clicking within the class to be selected. This enables analysis of the spatial and non-spatial distributions of each continuous variable for each category of the categorical variable (Theus and Urbanek 2009). Similarly, classes for continuous variables can be selected and the spatial and non-spatial distributions of the categories and other continuous variables can be compared. The height of each legend class changes in response to selections so that the coloured part of the legend class is proportional to the number of foreground objects (see figure 4.3).

The legend classes can also be brushed with the mouse. Brushing each class displays a pop-up label containing the number of selected and unselected objects in that class. Hidden objects are excluded (see the explanation of ternary selection in section 4.1).

4.3 Interactive parallel coordinates plot

A PCP is visualisation technique to represent high-dimensional data sets on a two-dimensional surface (Inselberg 2002). Variables are represented by a series of parallel axes with either a vertical or horizontal orientation. Objects in the data set are represented by continuous lines that intersect each axis at a point corresponding to the value of the associated variable. The PCP is best suited to represent continuous data although the lines can also be coloured according to the value of an additional variable which may be either categorical or continuous. The PCP thus provides both an overview of multivariate relationships in the data as well as retaining details of each individual object in the data.

The relationship between the categories and the predictor variables is explored with the interactive PCP shown in figure 4.5. The lines in this PCP are coloured blue if the associated county was won by George Bush and red if the associated county was won by John Kerry. This colour scheme is inherited from the legend in figure 4.2(a).

A difficulty with PCPs is over-plotting when the number of objects is large.

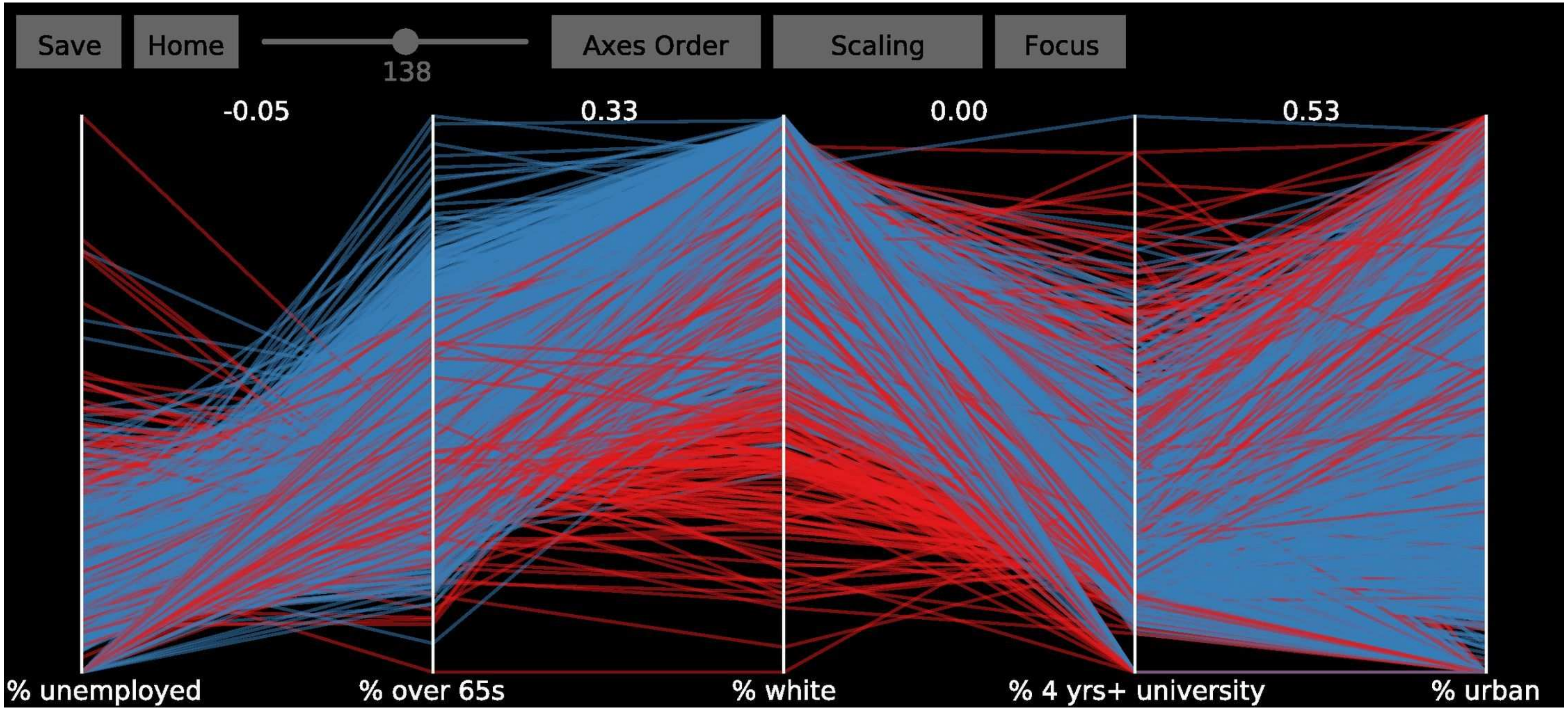


Figure 4.5: Visualisation of the relationship between the actual categories and the GWDA predictor variables in attribute space using a PCP. The categories are the 2004 US presidential election results and there are five predictor variables. Each predictor variable is represented by a single vertical axis in the and the lines are coloured by category so that so that counties won by George Bush are coloured blue and counties won by John Kerry are coloured red (see figure 4.2(a)).

The PCP in figure 4.5 implements two ways to reduce visual clutter:

1. The axes are automatically ordered so that the sum of the correlations between adjacent axes is maximized. This means that highly correlated variables tend to be clustered which minimizes the number of 'line crossings'. The heuristic OSL2 (Ordering Single Link) algorithm suggested by Hurley (2004) is used to automatically order the axes. It is also possible for the user to manually change the axes position using the mouse since the OSL2 algorithm is not guaranteed to identify the optimal solution.

2. The opacity of the lines can be adjusted using a slider in the PCP menu bar. This makes it easier to contrast areas of high line density with areas of low line density (Theus 2008).

The scale for the axes in the interactive PCP shown in figure 4.5 can also be changed, as suggested by Andrienko and Andrienko (2001). A 'Scaling' drop-down list in the PCP menu bar allows three different scalings to be applied:

1. Minimum-maximum scaling. Each variable is scaled linearly from the minimum to the maximum point on each axis (this is the default). The PCP in figure 4.5 has this scaling.

2. Scaling by median and quartiles. The median of each variable is mapped to the mid-point of each axis and the first and third quartiles mapped to the same horizontal lines for all variables. The remaining values are found by linear interpolation. The PCP in figure 4.11(b) has this scaling. The median for each variable is represented by a yellow circle superimposed on the associated axis. The position of the first and third quartiles is marked by a short yellow horizontal bar and the inter-quartile range, by a vertical yellow bar.

3. Scaling by mean and standard deviation. The mean of each variable, $\bar{x}$ is mapped to the mid-point of each axis and $\bar{x} - \sigma$ and $\bar{x} + \sigma$ are mapped to

the same horizontal line for all variables. The remaining values are found by linear interpolation.

Objects within each category (election outcome) can be selected by mouse-clicking in the interactive categorical legend (figure 4.2(a)). Lines for unselected objects are coloured grey in the PCP (see figure 4.11(b)).

Since GWDA involves inversion of the covariance matrix, high correlations between pairs of predictor variables can result in very small determinants and hence un-invertible matrices. Additionally, the inclusion of two highly correlated variables results in redundant information in the model and one of them should therefore be excluded from the analysis. Scatter plot matrices are commonly used to visually identify correlations between pairs of predictor variables but they occupy a lot of screen space for even a moderate number of variables. This thesis suggests using the PCP to identify highly correlated variables as follows:

1. The correlations between pairs of predictor variables are displayed at the top of the PCP between the associated axes.

2. The PCP axes are automatically ordered so that highly correlated variables are placed next to each other, as described above.

3. The order of the PCP axes can also be changed by the user using the mouse to explore the effect of different orderings. Wegman (1990) showed that for an m dimensional PCP, $\frac{m+1}{2}$ re-orderings are required to see all adjacencies between the variables.

Although a multicollinearity analysis should be done prior to classification, this research proposes using GWDA for exploratory purposes and therefore does not focus on this issue. If the results indicate multicollinearity, there are two options. Either one of the highly correlated variables should be omitted from the model or alternatively, the first few few principal components of the predictor variables should be used as predictor variables for the model. Brunsdon et al. (2007) take the latter approach in their case study.

4.4 Fluctuation diagram

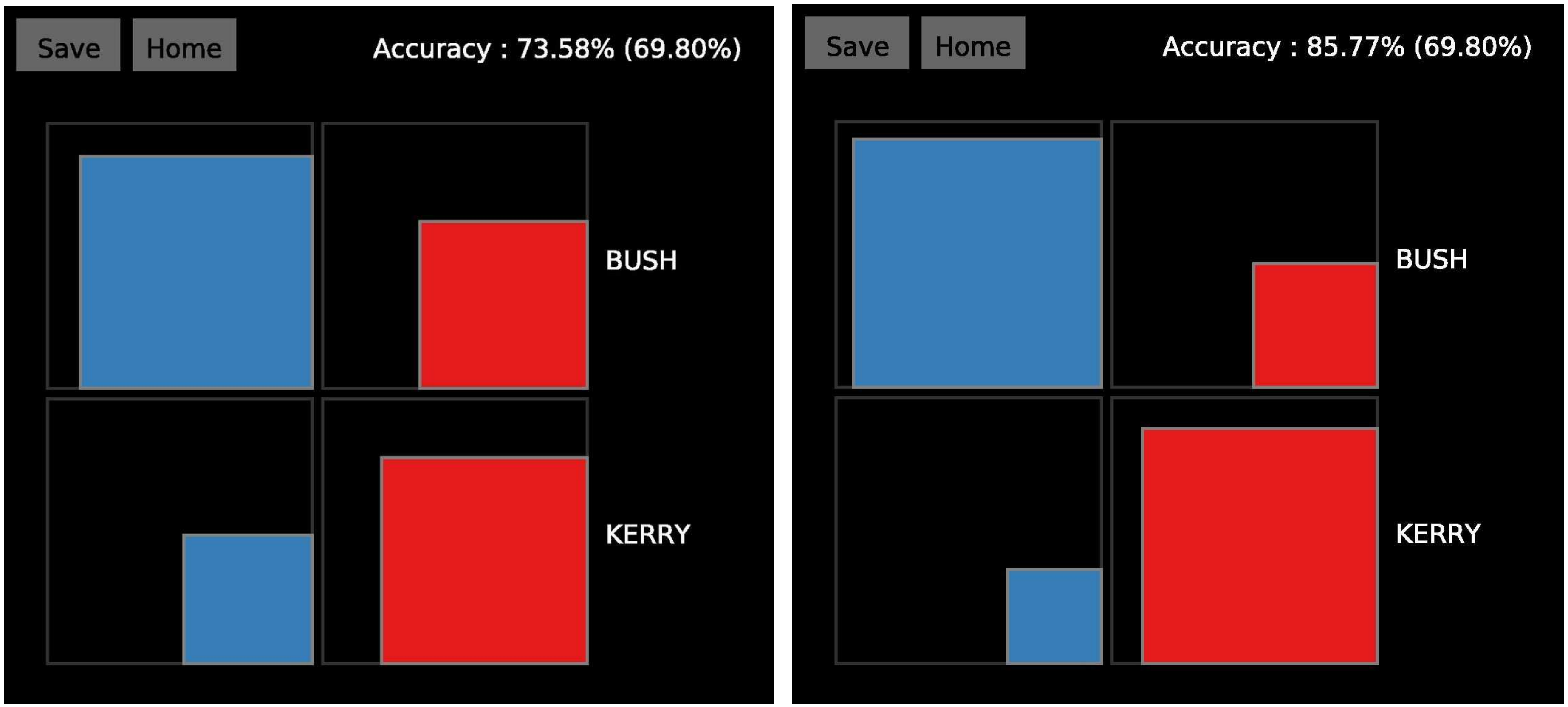
In remote sensing, a well-known method used to assess the classification accuracy is the confusion matrix (Afifi et al. 2004). This research uses graphical representations of the LDA and GWDA confusion matrices to assess their classification accuracies.

The confusion matrix works as follows. If there are k categories, a $k \times k$ matrix is required to tabulate all the possible combinations of *actual* and *assigned* categories. The cell in row i and column j, $c_{i,j}$ of the confusion matrix contains the number of objects assigned to category i but belonging to category j. Each diagonal cell contains the number of correctly classified objects in the corresponding category. Each off-diagonal cell contains the number of misclassified objects for that combination of i and j. Summing the rows of the confusion matrix ($\sum_i c_{i,j}$) gives the number of objects belonging to each category. Summing the columns of the confusion matrix ($\sum_j c_{i,j}$) gives the number of objects assigned to each category.

Fluctuation diagrams are commonly used in matrix visualisation (Hofmann 2008). In a fluctuation diagram, a $r \times c$ matrix is represented by a fluctuation diagram with $r \times c$ tiles so that the area of each tile is proportional to the corresponding matrix cell value. Therefore, fluctuation diagrams are used to represent confusion matrices in this research.

Interactive fluctuation diagrams were developed to visualise and explore the LDA and GWDA confusion matrices. Figure 4.6(a) shows the LDA confusion matrix and figure 4.6(b) shows the GWDA confusion matrix from the case study data set using fluctuation diagrams (see section 5.2.2). Since there are only two possible election outcomes for each county, Bush or Kerry, both confusion matrices are of size 2×2.

In this implementation, the following novel features have been added to a fluctuation diagram to make it suitable for exploring the LDA and GWDA confusion matrices:



(a) LDA confusion matrix.

(b) GWDA confusion matrix.

Figure 4.6: Visualisation of the LDA and GWDA confusion matrices using interactive fluctuation diagrams. The columns of the fluctuation diagrams are coloured according to the legend in figure 4.2(a) so that counties won by George Bush are coloured blue and counties won by John Kerry are coloured red.

- The area of each fluctuation diagram cell is proportional to the number of objects in the confusion matrix cell, divided by the number of objects in the corresponding category. This method highlights cells with a high classification/misclassification accuracy rather than simply cells with a large number of objects.

- The columns of the fluctuation diagram are ordered by classification accuracy from left to right, in descending order.

- The fluctuation diagram cells are coloured according to their actual category. The columns of the fluctuation diagrams in figures 4.6(a) and 4.6(b) are coloured according to the legend in figure 4.2(a). This is a useful perceptual cue to indicate the *actual* category that each cell represents. The assigned category labels are shown at the end of each row. If there are even a moderate number of categories, the category names would overlap if displayed at the top of each column. Therefore, only one column name is displayed at a time and this appears when cells are brushed with the mouse.

- Interactivity is permitted through selection and brushing of each cell in the fluctuation diagram. Brushing each cell displays a pop-up label containing the expected number of objects assigned to that cell and the assigned number of objects. All objects in a cell can also be selected by clicking with the mouse within the cell.

- The actual classification accuracy is shown in the upper right corner of the fluctuation diagram (first percentage). The figure in parentheses is the classification accuracy if the classification was based on chance. It is computed using the observed category frequencies as prior probabilities. If p_j is the prior probability that an object belongs to category j (see section 2.3.2), then ${p_j}^2$ is the probability that an object belonging to category j is assigned to category j. $\sum_j {p_j}^2$ is therefore the total probability of correct classification. This figure provides a benchmark to interpret the actual classification accuracy (Afifi et al. 2004).

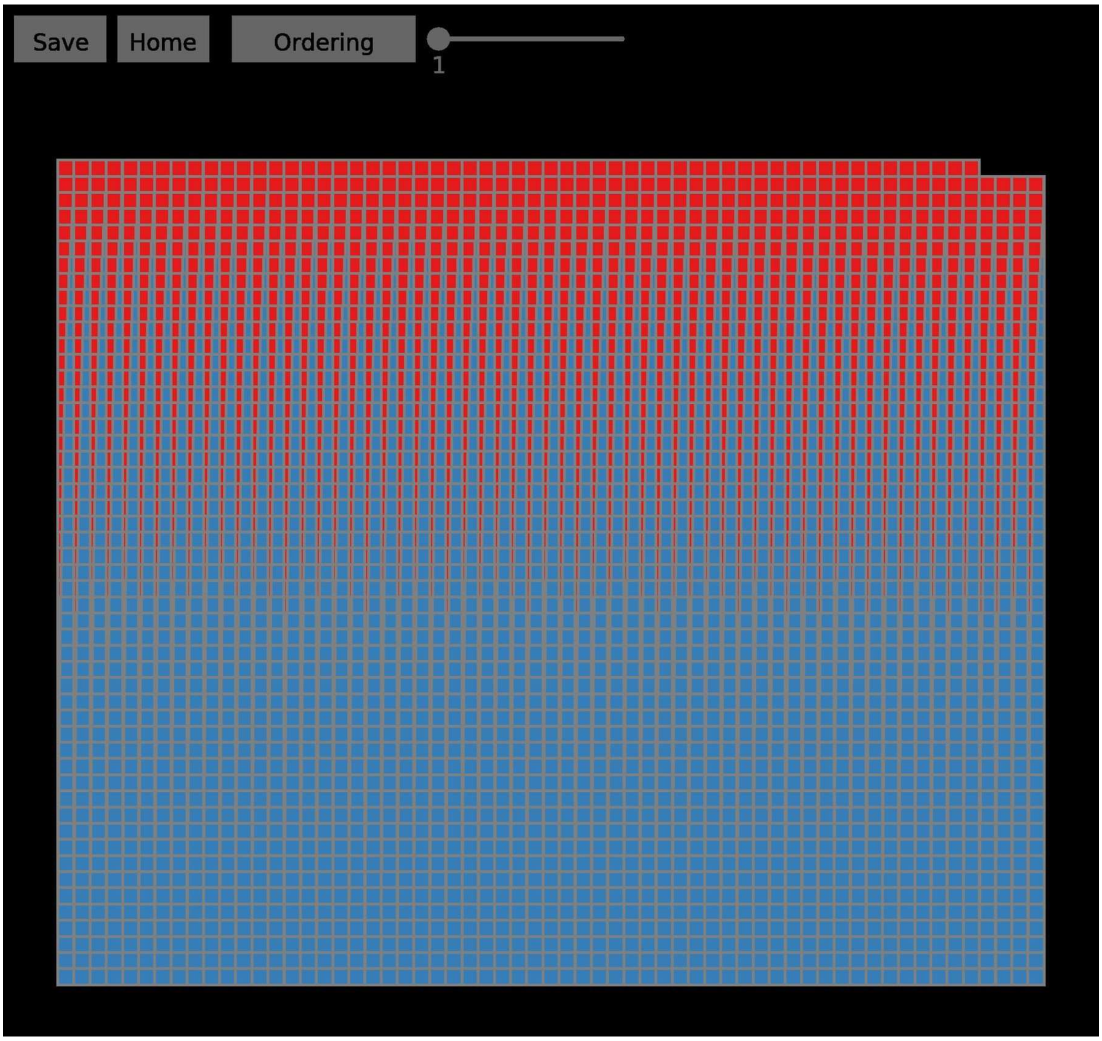
4.5 Posterior probability treemap

A treemap is a method to visualise hierarchical data using a recursive partitioning of space (Urbanek 2008). Each node in the hierarchical data set is represented by a rectangle. Starting with the root node, each rectangle is partitioned so that the area of each partition is proportional to the value the corresponding child node. Partitioning ends when there are no further child nodes — these are referred to as leaves.

In discriminant analysis, each object in a k category data set is characterised by k posterior probabilities. Although they do not form a traditional hierarchical data set in the manner described above, the posterior probabilities for each object represent a hierarchy with a single level. This suggests a treemap approach might be appropriate. Figure 4.7(a) shows a treemap visualisation an example of the GWDA posterior probabilities after GWDA was applied to the case study data set (see section 5.2.2).

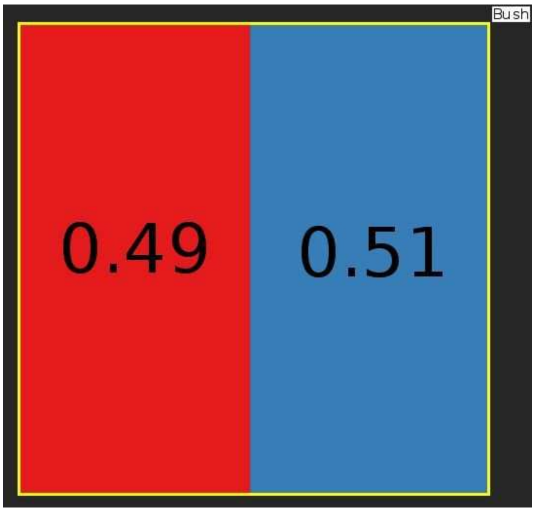
The treemap implementation used in this research has only one layer in the hierarchy. It works as follows:

Each object is represented by a square which is sub-divided into k sub-rectangles so that the area of the j^{th} sub-rectangle is proportional to the j^{th} posterior probability (there are k categories). The sub-rectangles are assigned colours corresponding to the category each posterior probability refers to (see the legend in figure 4.2(a)). The objects are ordered from bottom left to top right according to the posterior probability of a particular category which is chosen using the 'Ordering' drop-down list in the menu bar of the treemap. The slider on the menu bar allows adjustment of the scale at which brushed objects are drawn. This is useful when there is a large number of objects. When the scale is greater than one, the square representing the brushed object is magnified and appears slightly transparent to preserve context.



(a) Treemap visualisation of GWDA posterior probabilities for 3107 US counties. Each county is represented by a cell which is subdivided in proportion to the posterior probabilities for that county.

Figure 4.7: Visualisation of the GWDA posterior probabilities for the 2004 US presidential results using an interactive treemap. The treemap cells are coloured according to the legend in figure 4.2(a) so that the areas representing the posterior probabilities for George Bush are coloured blue and areas representing the posterior probabilities for John Kerry are coloured red.



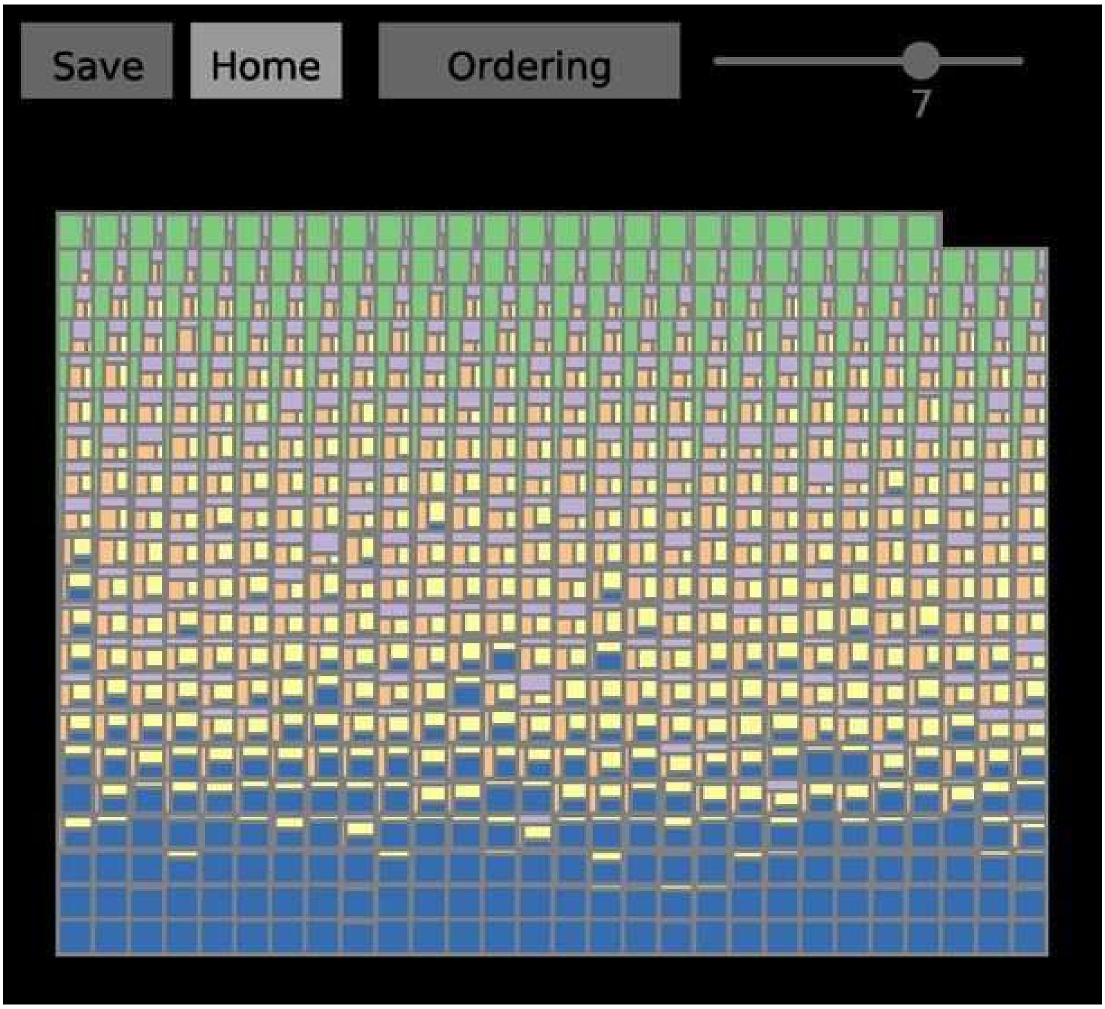
(b) A county (Franklin, Massachusetts) from figure 4.7(a) that was classified with a high level of uncertainty

Figure 4.7: Visualisation of the GWDA posterior probabilities using an interactive treemap. The treemap cells are coloured according to the legend in figure 4.2(a) so that the areas representing the posterior probabilities for George Bush are coloured blue and areas representing the posterior probabilities for John Kerry are coloured red.

The posterior probability treemap is used to explore uncertainty in the classification and to identify objects classified with a low degree of confidence, such as the object in figure 4.7(b). Because the visualisations in this geovisual analytics system shown in figure 4.1 are linked, the attributes of these objects can be explored. This enables the identification of clusters of similar objects in attribute and geographic space.

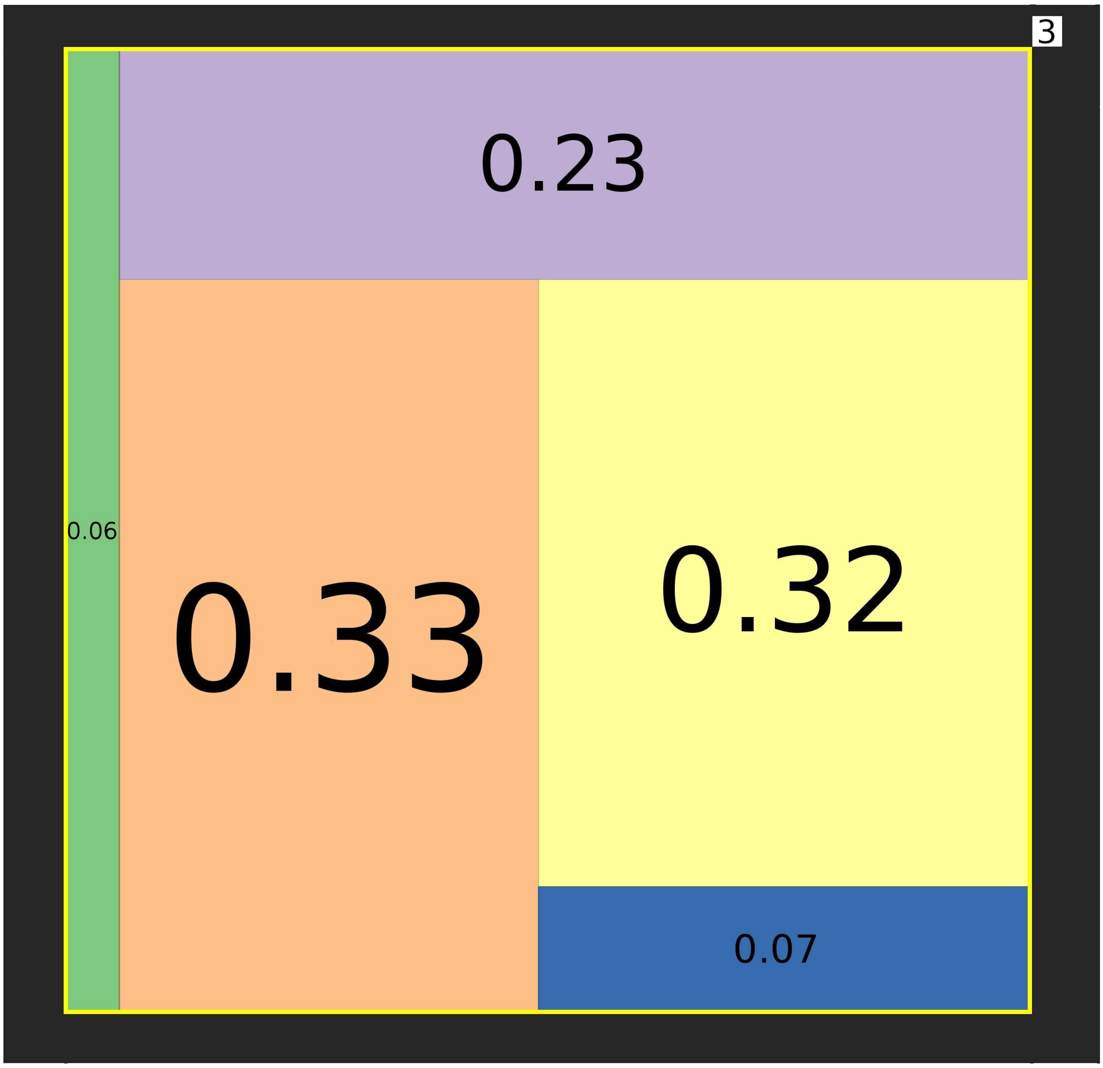
In figure 4.7(a) there are two categories and each square is broken into two sub-rectangles. The squares are ordered according to the posterior probability for the first category. Figure 4.8 shows a treemap visualisation of the GWDA posterior probabilities for a data set with five categories and 585 objects. This shows the nature of the partitioning more clearly than a data set with only two categories.

Brushing each cell in the treemap highlights it in yellow and displays a pop-up label showing the category that object is assigned to. When there is sufficient room, either after zooming or by adjusting the scale slider, the actual posterior



(a) Treemap visualisation of the GWDA posterior probabilities for five categories and 585 cells.

Figure 4.8: Visualisation of the GWDA posterior probabilities for a five category dataset using an interactive treemap.



(b) A cell from figure 4.8(a) that was classified with a high level of uncertainty.

Figure 4.8: Visualisation of the GWDA posterior probabilities for a five category dataset using an interactive treemap.

probability values are drawn in the centre of each sub-rectangle for the brushed object (see figures 4.7(b) and 4.8(b)).

4.6 Assessing the classification quality

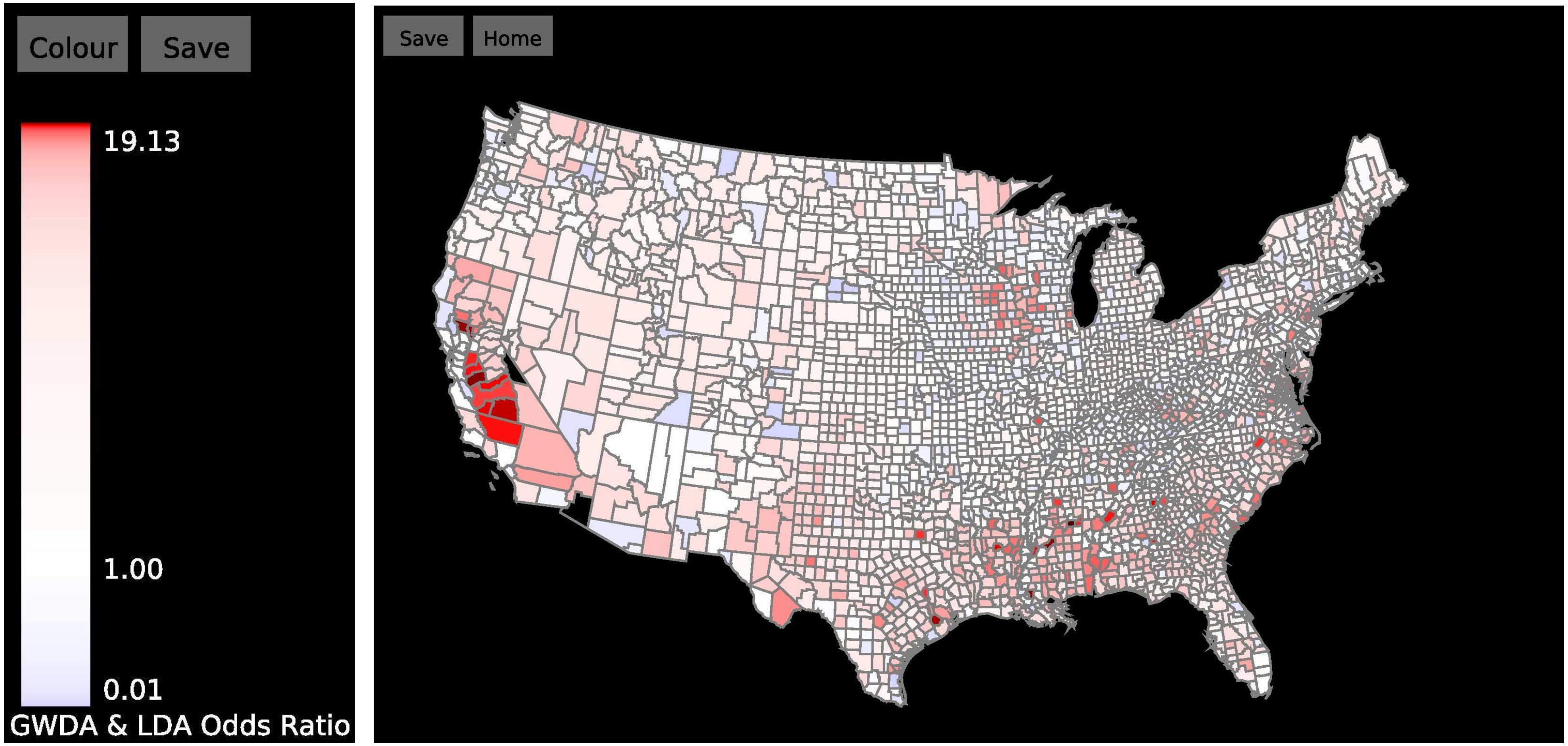
The quality of the classification provided by LDA and GWDA can be compared in three ways:

- Comparison of the classification accuracies of GWDA and LDA is possible using two fluctuation diagrams side by side (see figures 4.6(a) and 4.6(b)).

- The classification accuracies of GWDA and LDA will be very similar or identical if there is minimal non-stationarity. A more subtle approach is to compute the odds ratio of the GWDA and LDA posterior probabilities for membership of the actual categories. This is a simple ratio where the GWDA posterior probability is the dividend and the LDA posterior probability is the divisor. If the odds ratio is equal to one, both methods predict membership of the actual category with the same degree of confidence. If the odds ratio is greater than one, GWDA predicts membership of the actual category with greater confidence. If the odds ratio is less than one, LDA predicts membership of the actual category with greater confidence. Figure 4.9 compares the performance of GWDA and LDA using this methodology.

  Figure 4.9(a) is an interactive legend showing the odds ratio values assigned to each object in the data set. A blue-white-red diverging colour scheme is used so that odds ratio values greater than one are assigned increasingly darker shades of red and odds less than one are assigned increasingly darker shades of blue. Odds ratios of one are assigned white. Figure 4.9(b) is a thematic map showing the spatial distribution of the odds ratio values.

- The GWDA and LDA assigned categories can also be mapped and compared to identify geographical variations between the two models (see figures 4.4(b) and 4.4(c)).



(a) Odds ratio legend.

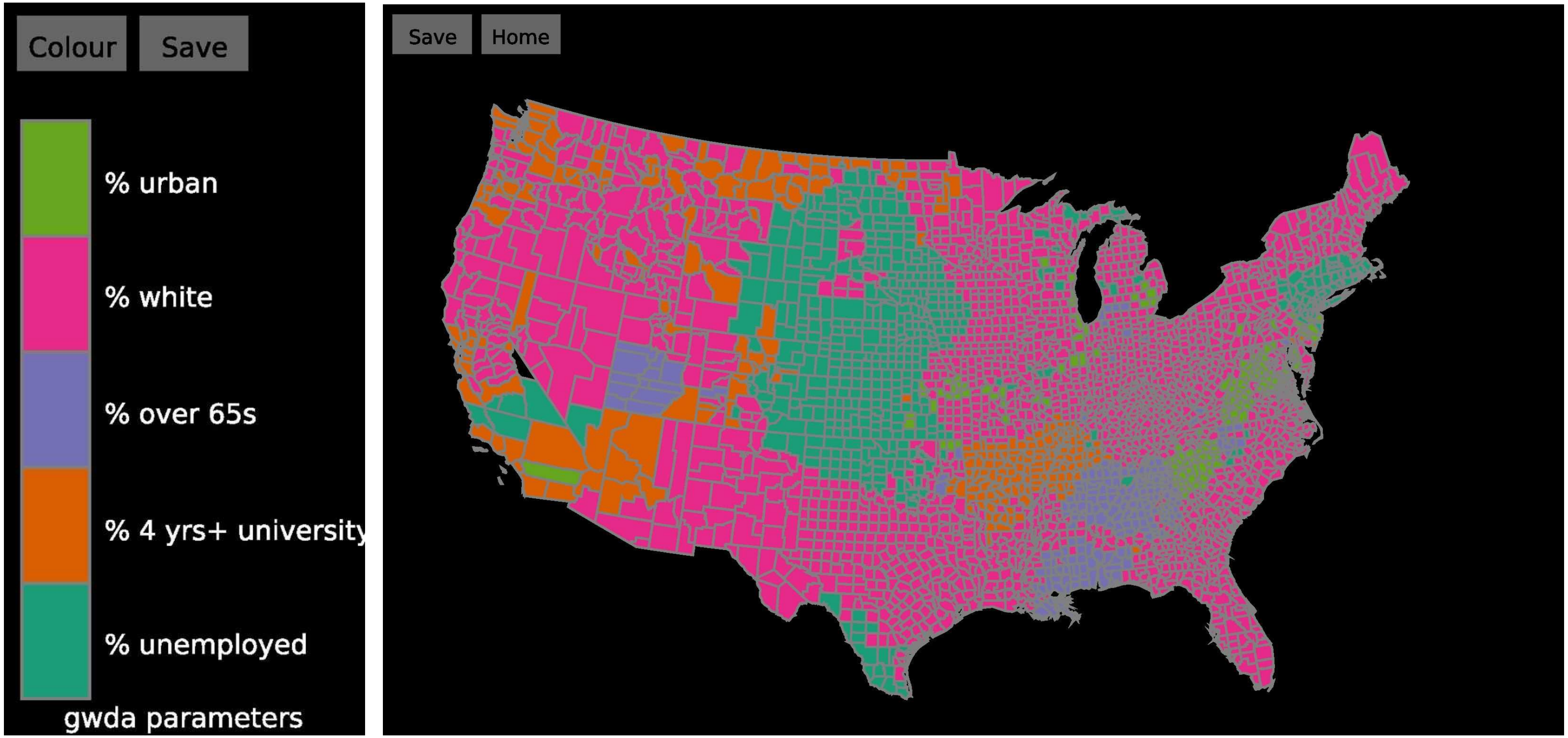
(b) Thematic map showing the spatial distribution of the odds-ratio values and coloured according to the legend in figure 4.9(a).

Figure 4.9: Visual comparison of GWDA and LDA posterior probabilities using an odds ratio. Counties coloured red are counties where the GWDA posterior probability for the actual election outcome is greater than the corresponding LDA posterior probability. Counties coloured blue are counties where the GWDA posterior probability for the actual election outcome is less than the corresponding LDA posterior probability. Counties coloured white are counties where the GWDA and LDA posterior probabilities are equal.

4.7 Spatial non-stationarity in the classification functions

GWDA models spatial non-stationarity in the relationship between the categories and the predictor variables through the spatially varying classification function parameters (see section 2.5.2). If the predictor variables are standardised prior to the classification, the parameters measure the relative contribution each variable makes towards the corresponding classification function. The extent and nature of spatial non-stationarity in the GWDA classification functions is explored by taking the winning classification function for each geographical area and comparing the parameter values for each variable. The variables are also ordered by the absolute value of the parameters. This allows the most influential variable(s) to be determined, in terms of assigned category for each geographical area. Figure 4.10 shows the linked visualisation tools used to analyse spatial non-stationarity in the classification functions.

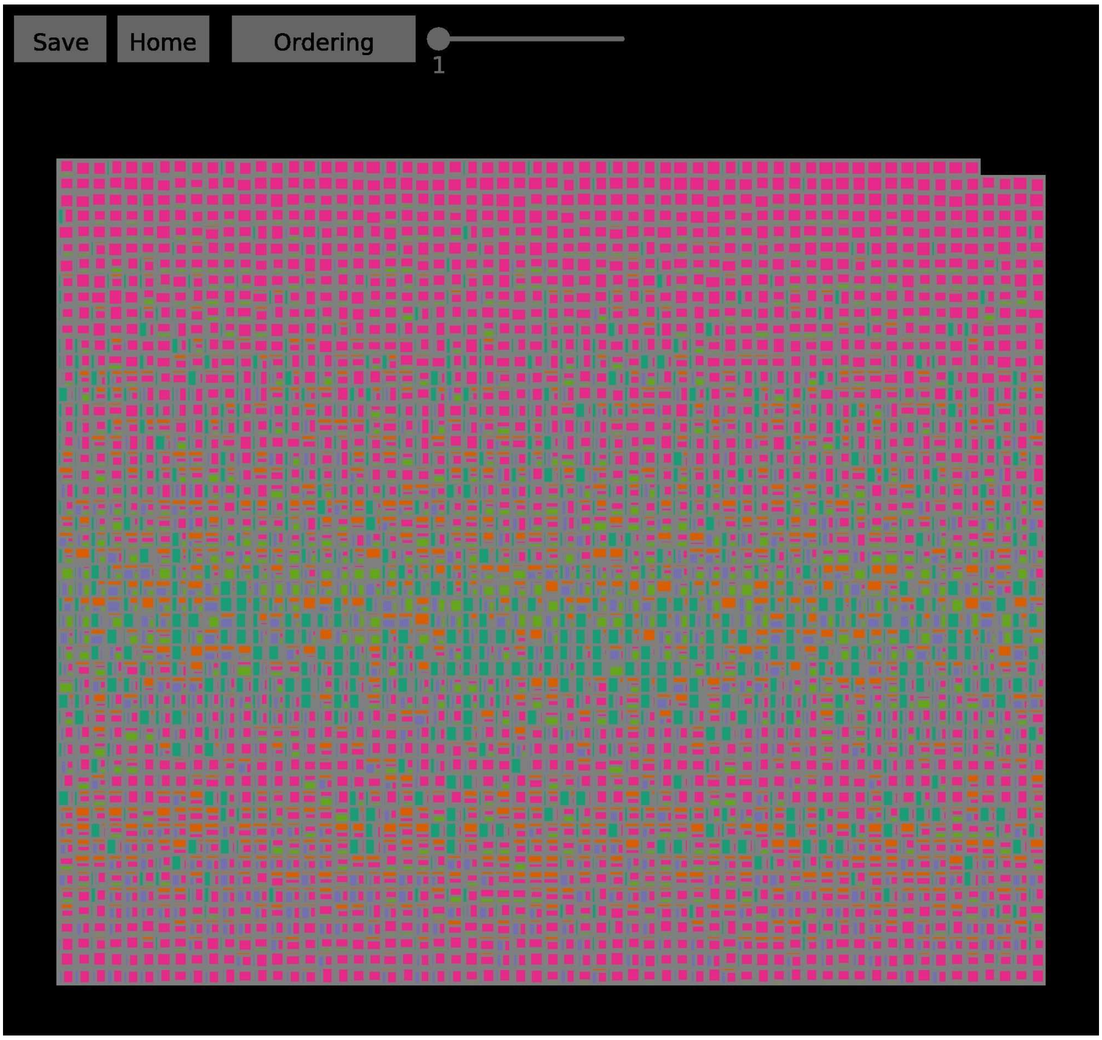
Figure 4.10(a) is an interactive categorical legend showing the colours assigned to each predictor variable and figure 4.10(b) is a thematic map showing the spatial distribution of the most important variable for each county. Figure 4.10(c) is a treemap visualisation showing the absolute parameter values for each county. This treemap is conceptually identical to that described in chapter 4.5 except that the area of each sub-rectangle is proportional to the absolute contribution of each variable to the winning classification function at each location instead of a posterior probability. The treemap is ordered by the contributions of the variable *% white* (pink) from most negative to most positive. Counties where the importance of this variable is strongly negative are shown at the bottom of the treemap. Counties where the importance of this variable is strongly positive are shown at the top of the treemap. The case study described in chapter 5 demonstrates how this representation enables a visual analysis of spatial non-stationarity.



(a) GWDA parameter legend.

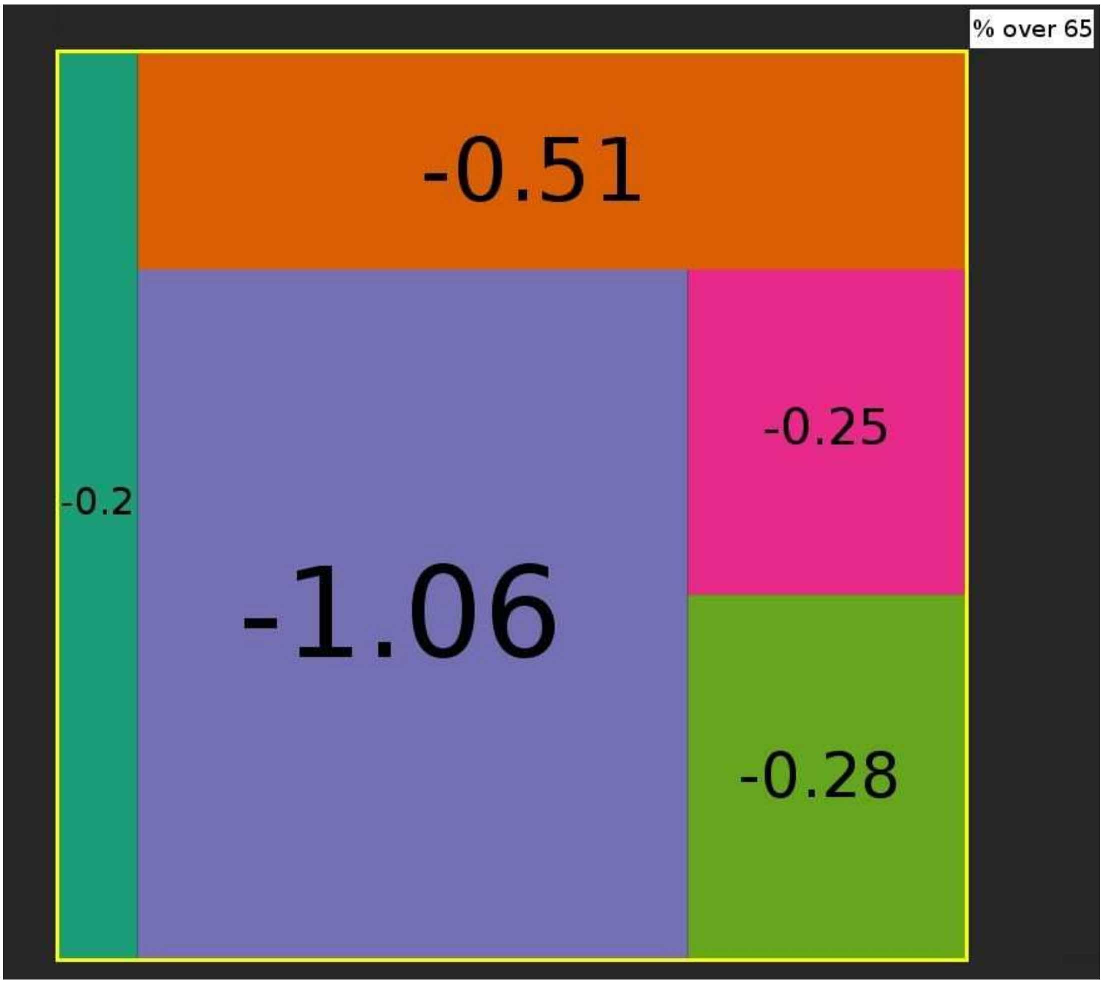
(b) Thematic map showing the spatial distribution of the most significant GWDA classification function parameters. This map is coloured according to the legend in figure 4.10(a).

Figure 4.10: Visualisation of spatial non-stationarity in the GWDA classification function parameters. The colours assigned to each parameter are shown in figure 4.10(a).



(c) Visualisation of the absolute GWDA classification function parameter values for each county using an interactive treemap. This treemap is coloured according to the legend in figure 4.10(a).

Figure 4.10: Visualisation of spatial non-stationarity in the GWDA classification function parameters. The colours assigned to each parameter are shown in figure 4.10(a).

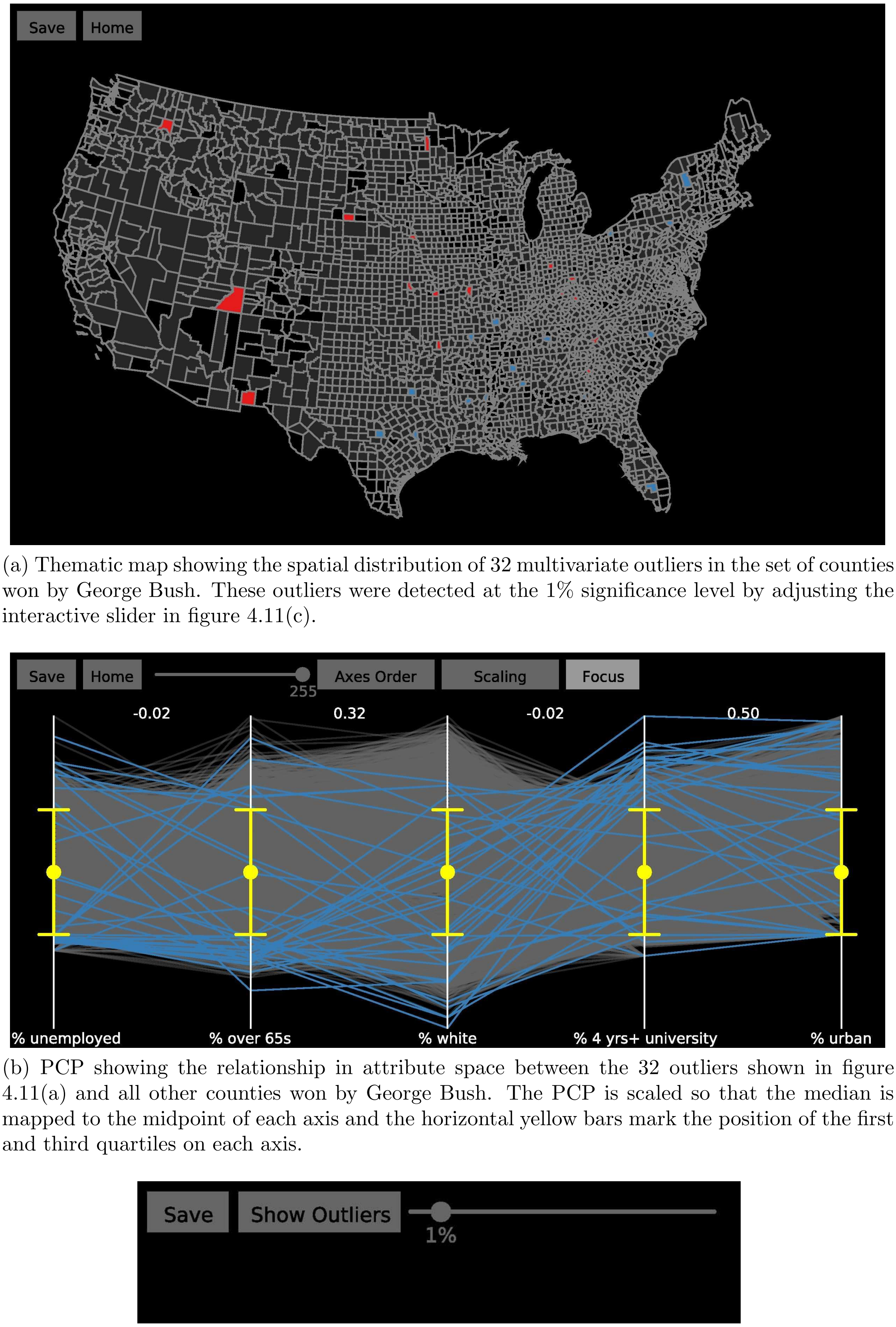


(d) A single cell from the interactive treemap in figure 4.10(c) showing the absolute GWDA classification function parameter values for a single county.

Figure 4.10: Visualisation of spatial non-stationarity in the GWDA classification function parameters. The colours assigned to each parameter are shown in figure 4.10(a).

4.8 Outliers

A method to automatically detect multivariate outliers within each category has also been implemented. The Mahalanobis distance squared from each observation to its category mean follows a chi-squared distribution with m degrees of freedom (Manly 2005) and this is used to identify outliers at different significance levels. Figure 4.11 shows 32 outliers detected at the 1% significance level for counties won by George Bush in a map (figure 4.11(a)) and a PCP (figure 4.11(b)). The PCP is scaled so that the median is mapped to the midpoint of each axis and the horizontal yellow bars mark the position of the second and third quartiles on each axis (see section 4.3). A slider is provided for users to adjust the significance level interactively (figure 4.11(c)).



(a) Thematic map showing the spatial distribution of 32 multivariate outliers in the set of counties won by George Bush. These outliers were detected at the 1% significance level by adjusting the interactive slider in figure 4.11(c).

(b) PCP showing the relationship in attribute space between the 32 outliers shown in figure 4.11(a) and all other counties won by George Bush. The PCP is scaled so that the median is mapped to the midpoint of each axis and the horizontal yellow bars mark the position of the first and third quartiles on each axis.

(c) Interactive slider to control the significance level of the outliers displayed in figures 4.11(a) and 4.11(b).

Figure 4.11: Visual analysis of multivariate outliers.

5 Case study

5.1 Introduction

This chapter describes a case study demonstrating the use of the geovisual analytic tools described in chapter 4 to investigate the performance of GWDA. The case study uses GWDA to model the relationship between the 2004 US presidential election results and five socio-economic indicators. Since GWDA won't necessarily work well with all categorical spatial data sets, the data requirements of the method are explained first. Then, the data set used for the case study is described. Next, the implementation of GWDA used for the case study is explained. Finally, the visualisation tools from chapter 4 are used to investigate the performance of the GWDA model in terms of classification accuracy, classification uncertainty and spatial non-stationarity.

5.2 Selecting a data set

5.2.1 Data requirements

As explained in section 1.1, GWDA can only be used for social science applications since non-stationarity does not occur in the physical sciences. This is because physical laws do not vary spatially and therefore any apparent variation in the relationship between the categories and the predictor variables must be accounted for by additional variables.

As explained in section 2.3.1, discriminant analysis can be used, either to quantify the relative importance of each predictor variable in determining the categories or for supervised classification. However, GWDA is much better suited to analytical/descriptive type applications rather than supervised classification and Brunsdon et al. (2007) take an analytical/descriptive approach in their case study. The reason why GWDA is not well suited to supervised classification is because geographical position is taken into account when calibrating the classification functions. Training data are required for supervised classification and these are used to calibrate the GWDA classification functions. Thus, the resulting GWDA model reflects the spatial distribution of the training categories. If the spatial distribution of the training categories does not reflect the spatial distribution of the actual categories, objects could be misclassified by the GWDA model. However, the spatial distribution of the actual categories may not be known in advance and this makes it difficult to choose "good" training data locations. Additionally, the number of training data objects tends to be small relative to the number of unclassified objects. For example, Serra et al. (2007) use a training data set of area 41.5ha to train a discriminant analysis classifier which is subsequently used to classify a data set of area $20,363$ha. It is difficult to approximate the spatial distribution of the actual categories with so few training data objects.

Although Brunsdon et al. (2007) used fixed and adaptive bandwidths in their case study, this research found that an adaptive bandwidth approach works better in most cases. This is due to the nature of discriminant analysis, which requires k classification functions to classify a k category data set (see section 2.3). If a fixed bandwidth is used it must be chosen so that data for each of the k categories is reachable from each geographical location. If the fixed bandwidth is too small, it will not be possible to calibrate classification functions for all k categories from some locations. If the fixed bandwidth is too big, the local variations that GWDA is intended to model could be missed. Choosing a fixed bandwidth under these circumstances is extremely difficult. If GWDA is to be used with a fixed bandwidth, the spatial distribution of the categories should be evenly mixed throughout the data.

An adaptive bandwidth works better since it is specified as the minimum number of objects in each category required for calibration of the corresponding classification function, rather than a fixed distance. However, the spatial distribution of the categories also affects the adaptive bandwidth. If the categories are not reasonably evenly spatially mixed, the resulting adaptive bandwidths can extend over the entire data set. The principle of local weighting is not applied and the GWDA results are virtually identical to the LDA results.

To summarise, GWDA is suited to modelling processes in the social sciences where the categories are reasonably evenly mixed spatially.

5.2.2 Case study data set: 2004 US presidential election results

This case study uses the sample data set from the Exploratory Spatio-Temporal Analysis Toolkit (ESTAT) Geovisualization toolkit, developed by researchers at the GeoVISTA Center, Pennsylvania State University. (`www.personal.psu.edu/users/a/c/acr181/election.html`, accessed on the 8th of November 2011). This data set contains the 2004 US presidential election results at county level together with a selection of socio-economic variables.

There were three candidates in the 2004 US presidential election: George Bush, John Kerry and Ralph Nader who failed to win any counties. Although figure 4.4(a) shows a definite spatial pattern to the election results, and George Bush won more than four times as many counties as John Kerry (2531 to 576), the categories are reasonable spatially mixed to some extent (see the data requirements described in section 5.2.1). The exception to this pattern is the mid-west where John Kerry failed to win any counties.

The case study in the original GWDA paper by Brunsdon et al. (2007), uses GWDA to explore the relationship between the outcome of the 2005 UK general election and the following six census variables:

1. The percentage of economically active males unemployed.

2. The percentage of the adult population with no qualifications.

3. The percentage of pensioners in the population.

4. The percentage of non-white people in the population.

5. The percentage of owner occupied households.

6. The percentage of lone-parent households.

This case study uses five socio-economic variables from the ESTAT sample data set as predictor variables and four of these are similar to the above. They are:

1. The percentage unemployed.
2. The percentage of adults over 25 with 4+ years of college education.
3. The percentage of persons over the age of 65.
4. The percentage white.
5. The percentage urban.

These first four variables are equivalent to the first four variables used in the study by Brunsdon et al. (2007). Since the ESTAT sample data set did not contain similar variables for the percentage of owner occupied households or the percentage of lone-parent households, these were omitted these from the analysis. A new variable, *% urban* was added following a suggestion from the developers of the ESTAT application.

Four counties: Clifton Forge, Virginia; Mono, California; South Boston, Virginia and Yellowstone National Park, Montana recorded an equal number or no votes for Bush and Kerry. These were omitted from the analysis leaving a total of $3,107$ counties.

Since the ESTAT shape file is in decimal degrees and the version of GWDA developed for this research requires the computation of Euclidean distances, the data set was projected with a distance preserving planar projection (US Contiguous Equidistant Conic) in ArcGIS 10 prior to the classification.

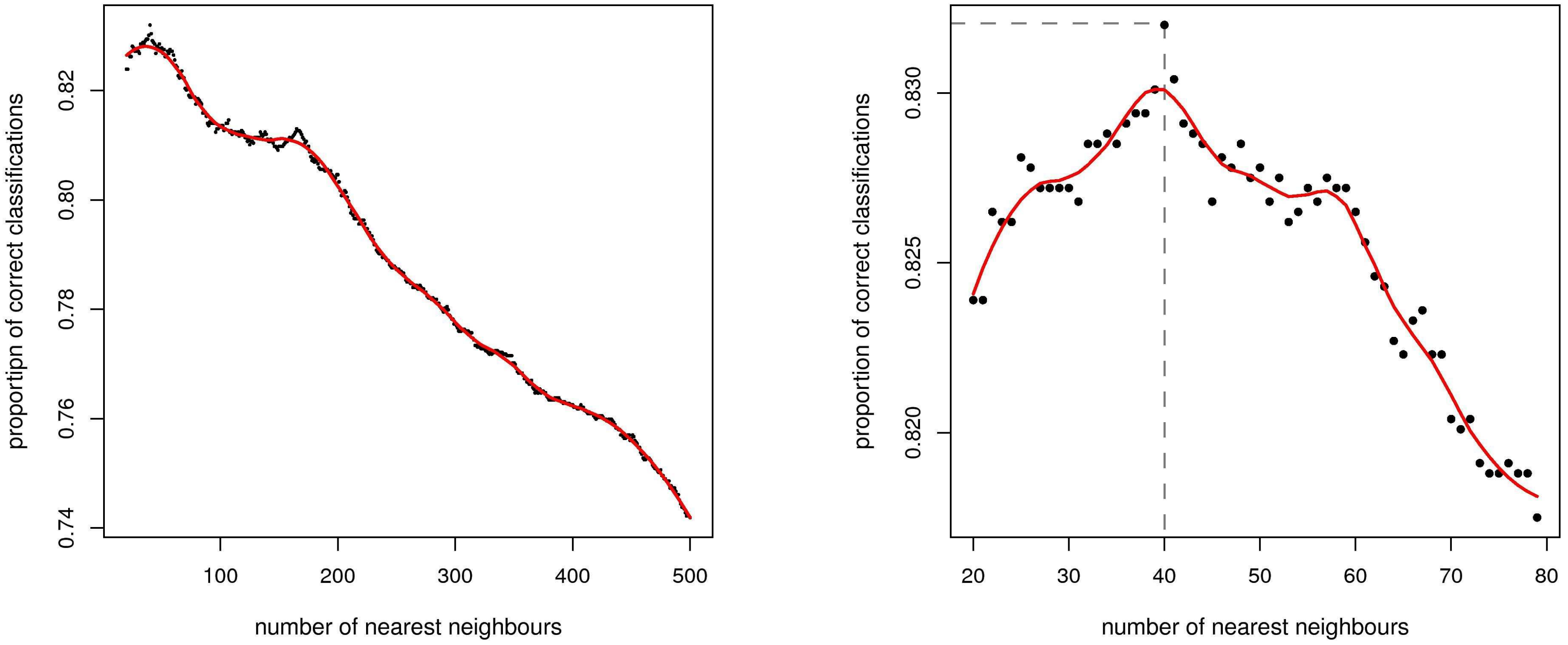
Although not all five US socio-economic variables are normally distributed, which is a theoretical requirement for GWDA (see section 2.4.1), this thesis uses GWDA as an exploratory method. In this context, the issue of multivariate normality is outside the scope of the research.

5.3 Implementation of geographically weighted discriminant analysis

GWDA was implemented as described in section 2.5.3. However, although the category means and covariance matrices were allowed to vary spatially, the prior probabilities were equal and fixed. This was for two reasons. Firstly, there was no reason to assume a natural bias in the 2004 US election results. Secondly, the discussion generated by the original GWDA case study of Brunsdon et al. (2007) concluded that the GWDA classification accuracy is improved by equal prior probabilities (Johnston and Pattie 2009; Brunsdon 2009), although the reasons for this are not clear.

The GWDA model was calibrated using an adaptive bandwidth (see section 5.2.1) and a bisquare kernel (see equation 2.10). Cross-validation was used to choose the optimum number of nearest neighbours by maximizing the proportion of correct classifications. 500 was chosen as a reasonable maximum number of neighbours since the total number of counties won by a single candidate was only 576 (John Kerry). The minimum number of neighbours was set at 20 since it was felt that calculations of the covariance matrices for fewer than 20 counties could be unstable. The results from the cross-validation are shown in figure 5.1 and the the optimum number of nearest neighbours was identified as 40 (see the object marked by the dashed line).

Following this, the GWDA model was calibrated for the same data set using an optimum number of nearest neighbours of 40 and a bisquare kernel. The five predictor variables were also standardised prior to classification so that the classification function parameters could be compared (see section 2.4.3). The results are discussed in the next section.



(a) The proportion of correct classifications for the range of nearest neighbours values $(20, 500)$.

(b) The proportion of correct classification for a range of nearest neighbours values $(20, 80)$. The optimum number of nearest neighbours is 40 which is emphasised by the dashed line.

Figure 5.1: Results from the cross-validation results to identify the optimum number of nearest neighbours for GWDA. The optimum number of nearest neighbours is difficult to see from figure 5.1(a), so figure 5.1(b) is a "close-up'.

5.4 Results

5.4.1 Introduction

The quality of the GWDA classification was assessed in terms of the classification accuracy and classification uncertainty. Then the classification function parameters were examined for evidence of non-stationarity using the methodology described in chapter 4.

5.4.2 Assessing the classification accuracy

The fluctuation diagrams described in section 4.4 were used to compare the classification accuracies of GWDA and LDA.

The LDA confusion matrix is shown in table 5.1 and visualised in the fluctuation diagram in figure 4.6(a).

Table 5.1: LDA confusion matrix

	Bush	**Kerry**	**Total**
Bush	1939	229	2168
Kerry	592	347	939
Total	2531	576	3107

The GWDA confusion matrix is shown in table 5.2 and visualised in the fluctuation diagram in figure 4.6(b).

Table 5.2: GWDA confusion matrix

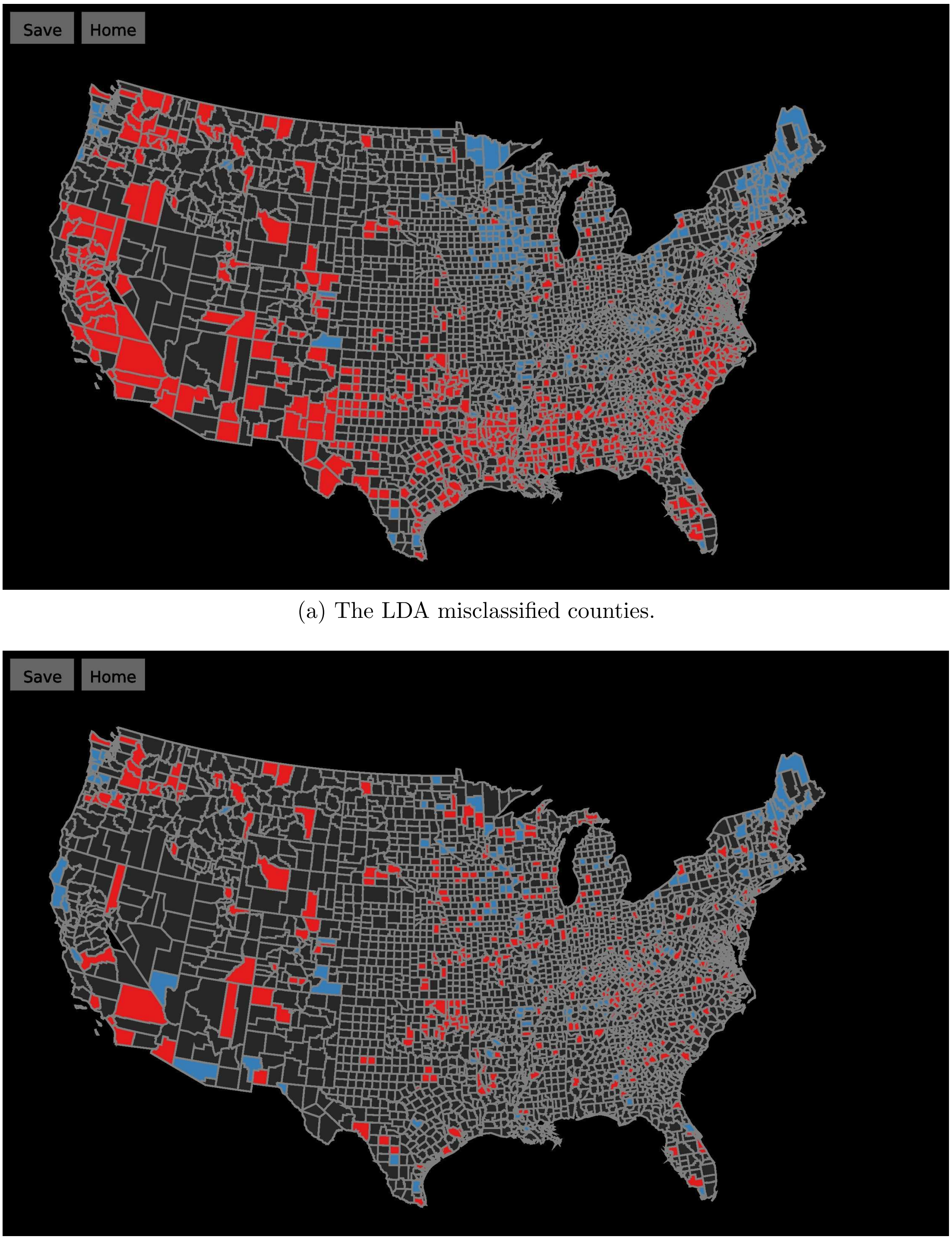
	Bush	**Kerry**	**Total**
Bush	2214	125	2339
Kerry	317	451	768
Total	2531	576	3107

The LDA classification accuracy is 73.58% (see figure 4.6(a)) and the GWDA classification accuracy is 85.77% (see figure 4.6(b)). If the data were classified by chance, using the observed categories frequencies as prior probabilities, the classification accuracy would be 69.8%. The GWDA classification accuracy is a significant improvement on both the LDA and chance classification accuracies. Brushing the fluctuation diagram cells in figure 4.6 revealed that the number of misclassified counties was 821 with LDA but only 442 with GWDA. The majority of this improvement is due to a reduction in the number of counties won by Bush but assigned to Kerry (592 with LDA versus 317 with GWDA).

Examination of the thematic maps in figure 4.4 reveal that the spatial pattern of the election results is also captured more accurately by GWDA. Comparison of the map of the actual election results (figure 4.4(a)) with the LDA assigned election results (figure 4.4(b)) reveals an interesting spatial pattern to the LDA misclassifications. Counties in California, the South and the East Coast tend to be assigned to Kerry by LDA while counties in the interior of the US tend to be assigned to Bush. For example, counties in New England and Minnesota were won by Kerry but assigned to Bush by LDA, while many counties in California and the South were won by Bush but assigned to Kerry. The spatial pattern of the GWDA assigned election outcome (figure 4.4(c)) does not display this trend and is a closer match to the actual results in figure 4.4(a).

One justification for the use of GWR is when strong spatial autocorrelation of the residuals is observed in traditional linear regression models (Fotheringham et al. 2002). If spatial autocorrelation of the residuals is much lower with GWR this is because geographically weighted models are better able to capture non-stationarity in the data. In the context of GWDA, the equivalent of regression residuals are misclassifications. Selection of the LDA and GWDA misclassified counties is possible by clicking with the mouse in the off-diagonal cells of the fluctuation diagrams in figure 4.6. Since the fluctuation diagrams are dynamically linked to the thematic maps the spatial patterns of the LDA and GWDA misclassifications can be seen in figure 5.2.

Figure 5.2(a) shows a clear pattern of spatial clustering in the LDA residuals. However, significantly less spatial clustering is evident in the GWDA residual map



(a) The LDA misclassified counties.

(b) The GWDA misclassified counties.

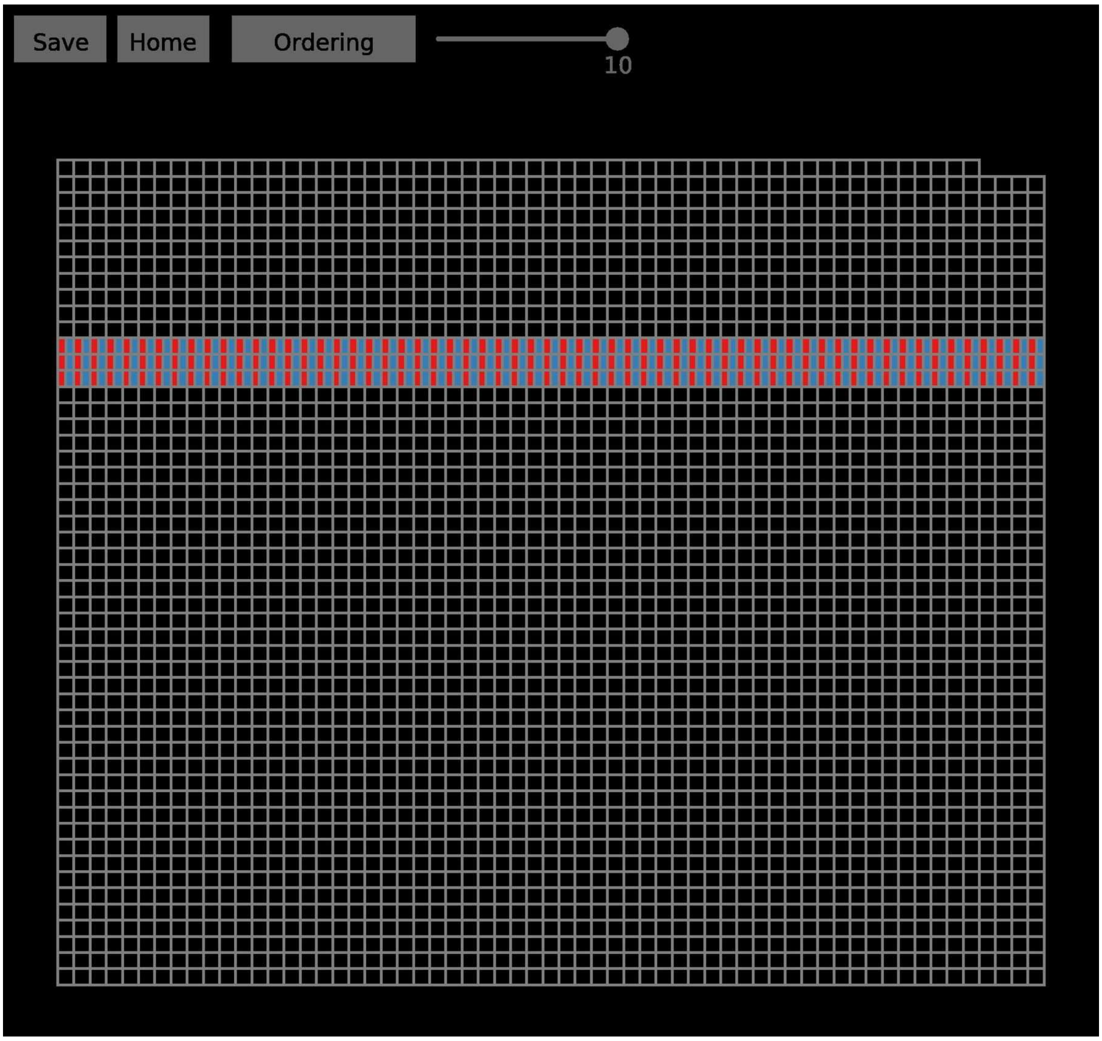
Figure 5.2: Thematic maps showing the spatial distribution of US counties misclassified by LDA and GWDA. The counties are coloured according to the legend in figure 4.2(a) so that counties assigned to George Bush are coloured blue and counties assigned to John Kerry are coloured red.

in figure 5.2(b). Following the logic of GWR, this indicates that the relationship between the election results and predictor variables is indeed non-stationary.

5.4.3 Assessing the classification uncertainty

The posterior probability treemap described in section 4.5 was used to explore uncertainty in the GWDA classification and further compare the performance of LDA and GWDA.

The treemap for the GWDA posterior probabilities is shown in figure 4.7(a). The cells are placed in ascending order from bottom left to top right according to the value of the posterior probability for Kerry which is coloured red. The posterior probability for Bush is shown in blue. These colours are identical to the colours used in the categorical legend for the election outcome in figure 4.2(a). Cells dominated by red towards the top of the treemap or blue towards the bottom represent counties classified with a low level of uncertainty. Some cells however were classified with a high level of uncertainty, such as Franklin, Massachusetts (figure 4.7(b)). Using the mouse, 183 counties classified by GWDA with a high level of uncertainty were selected from the middle of the posterior probability treemap (see figure 5.3(a)). These are counties where the posterior probabilities for a Bush or Kerry outcome are almost equal and range from 0.46 for Kerry and 0.54 for Bush in the lower left to 0.55 for Kerry and 0.45 for Bush in the upper right. Figure 5.3(b) shows the spatial distribution of these counties, coloured by actual election outcome. Two spatial patterns are evident. Firstly, there is general spatial clustering around counties near the great lakes. Secondly, the counties won by Kerry within this subset are mostly located in the east of the US. The linked LDA fluctuation diagram (figure 5.3(c)) and GWDA fluctuation diagram (figure 5.3(d)) reveal that GWDA does not result in an improved classification accuracy for both Kerry and Bush counties. For the counties won by Kerry, 52 counties were misclassified by LDA whereas 23 were misclassified by GWDA. However, for the counties won by Bush, 37 were misclassified by LDA whereas 57 were misclassified by GWDA.

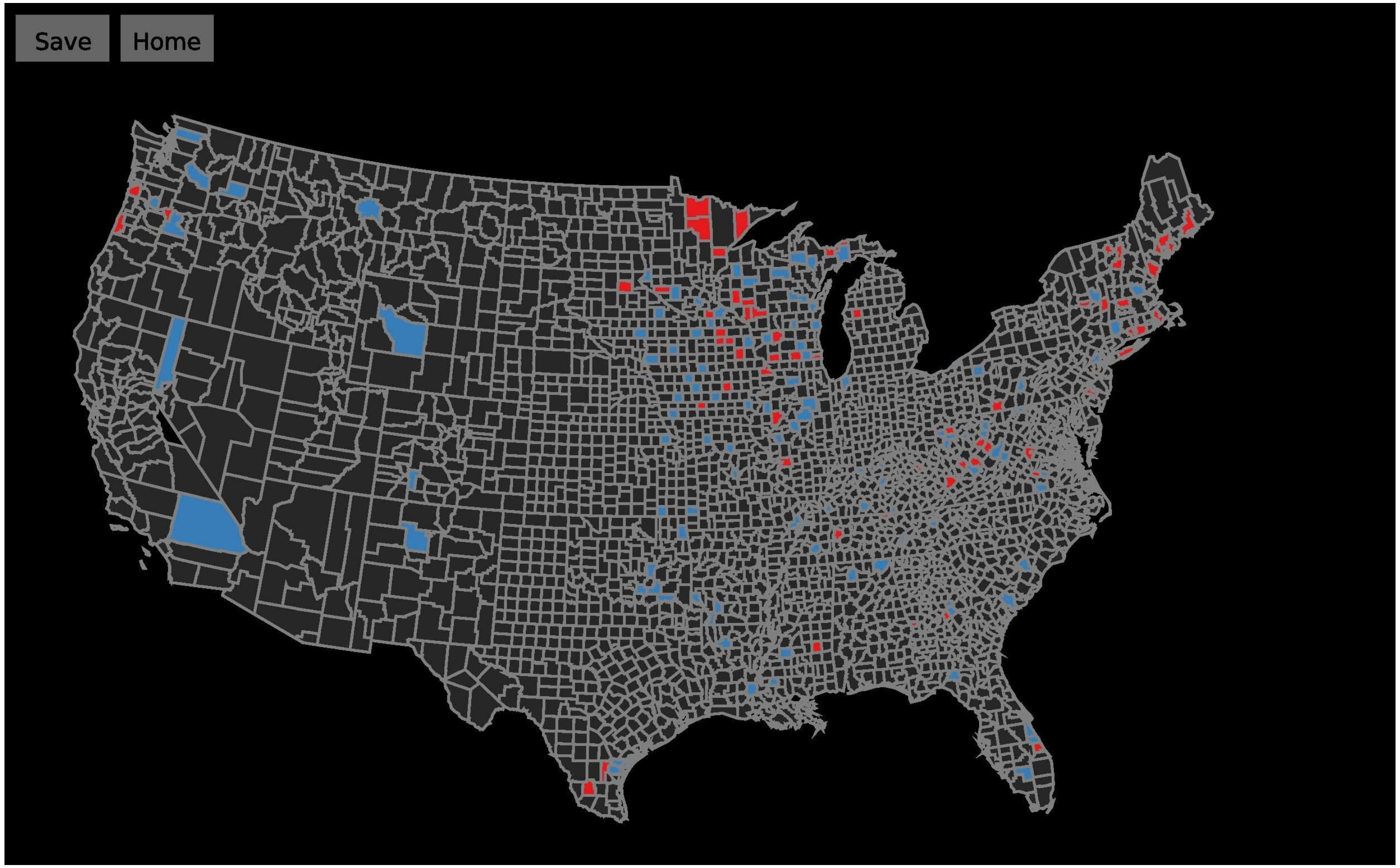


(a) 183 US counties classified by GWDA with a high level of uncertainty shown in an interactive treemap. These are counties with posterior probabilities very close to 0.5.

Figure 5.3: Visual analysis of 183 US counties classified by GWDA with a high level of uncertainty.

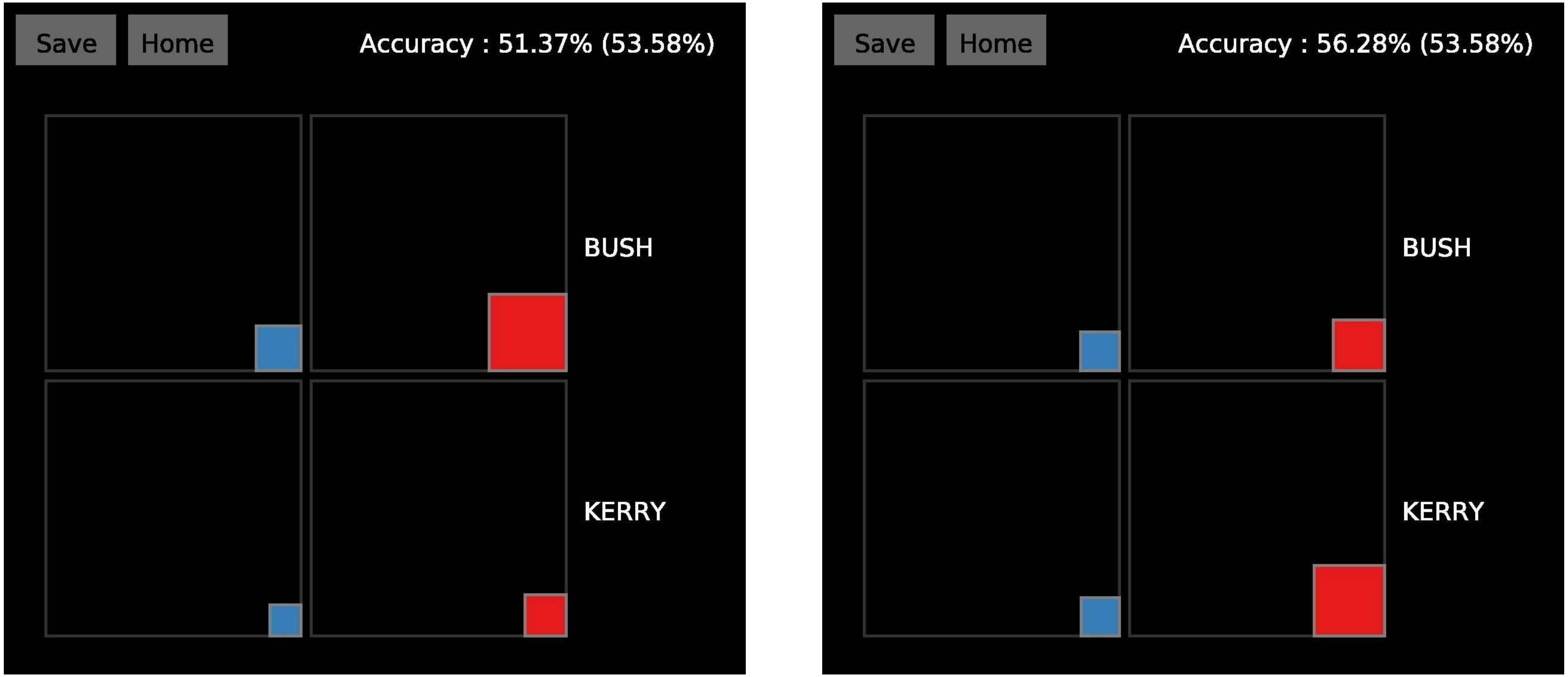
Figure 4.9(b) shows the spatial distribution of a comparison of the GWDA and LDA posterior probabilities using an odds ratio.

Counties coloured red in figure 4.9(b) are counties where the posterior probability for the GWDA assigned outcome is higher than the corresponding LDA posterior probability. Counties coloured blue are counties where the GWDA posterior probability for the actual category is lower than the corresponding LDA posterior probability. Most counties are almost white indicating that there is very little difference in the confidence in the prediction, in other words the odds



(b) Thematic map showing the spatial distribution of 183 US counties classified by GWDA with a high level of uncertainty. These counties are coloured according to the actual US presidential election outcome (see the legend in figure 4.2(a)).

Figure 5.3: Visual analysis of 183 US counties classified by GWDA with a high level of uncertainty.



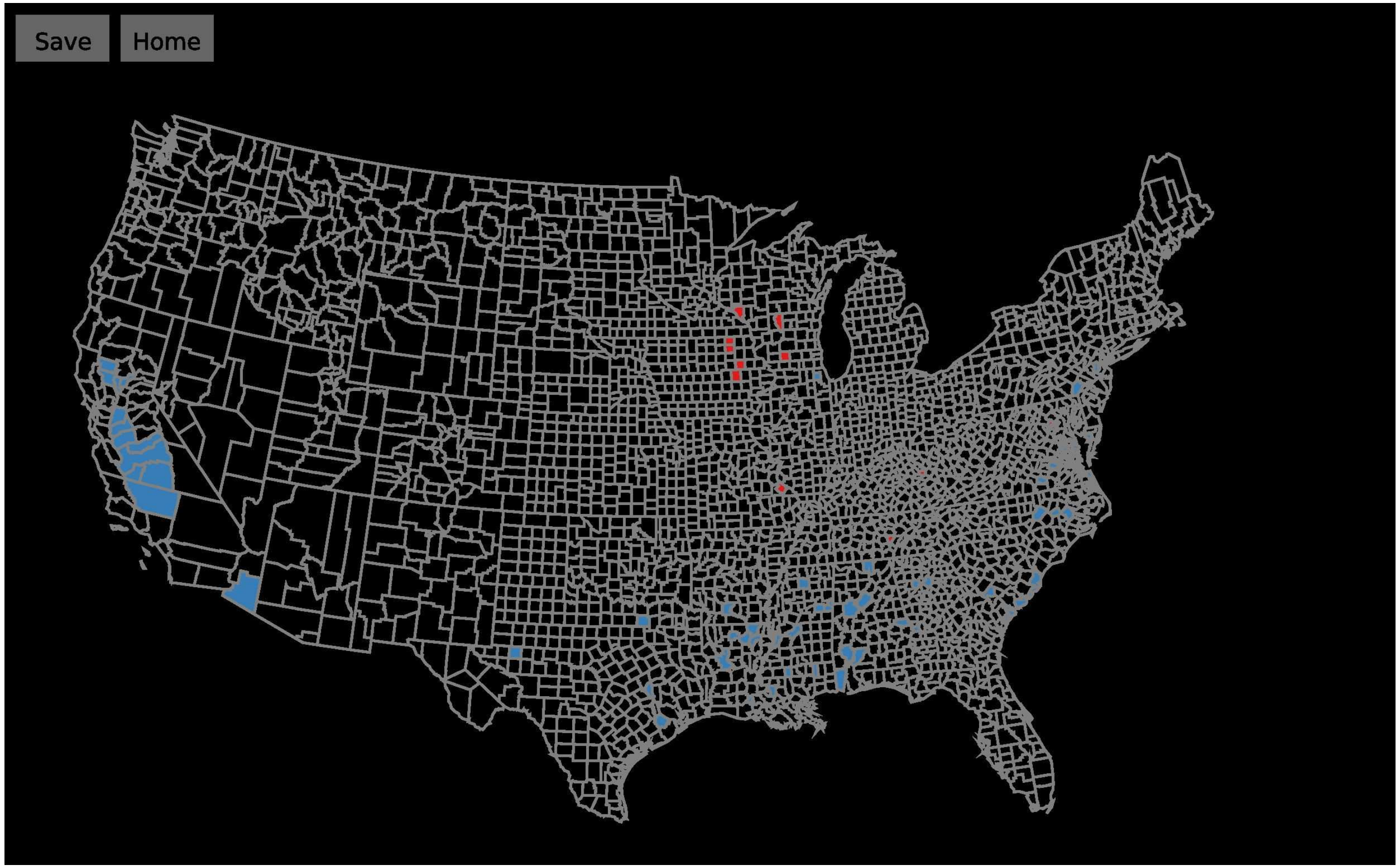
(c) Fluctuation diagram showing the LDA confusion matrix for 183 US counties classified by GWDA with a high level of uncertainty. The fluctuation diagram columns are coloured according to the legend in figure 4.2(a).

(d) Fluctuation diagram showing the GWDA confusion matrix for 183 US counties classified by GWDA with a high level of uncertainty. The fluctuation diagram columns are coloured according to the legend in figure 4.2(a).

Figure 5.3: Visual analysis of 183 US counties classified by GWDA with a high level of uncertainty.

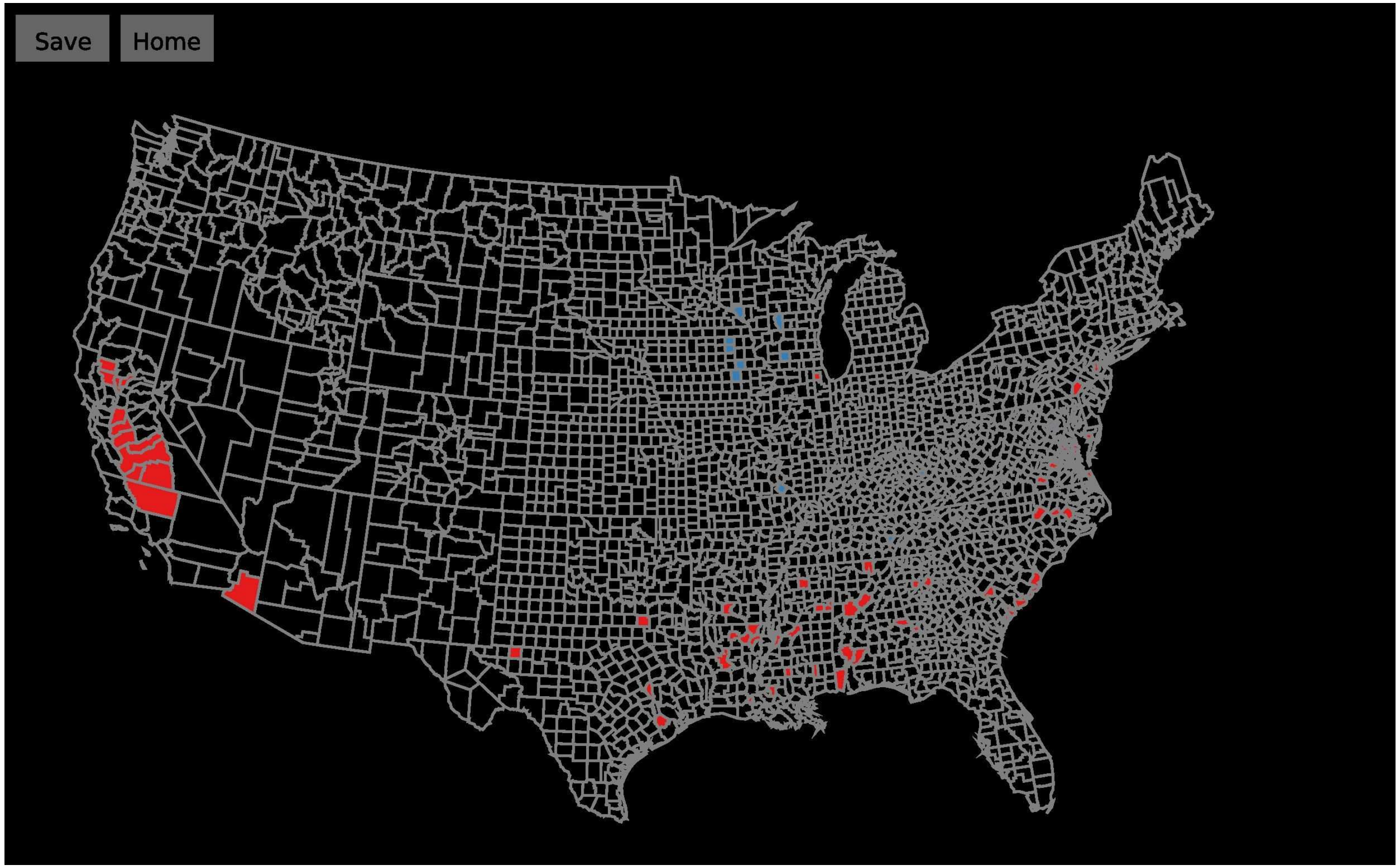
ratio is very close to one. Generally, the greatest reduction in uncertainty is seen in California and the South. Using the mouse, the 80 counties where GWDA resulted in the the greatest reduction in classification uncertainty were selected from the odds ratio legend in figure 4.9(a). The spatial distribution and classification accuracy for this subset is shown in figure 5.4.

The linked LDA and GWDA fluctuation diagrams (figures 5.4(d) and 5.4(e)) reveal that all of these counties were misclassified by LDA but only 7 were misclassified by GWDA. It is concluded that the reduction in classification uncertainty associated with the GWDA model was sufficient to improve the classification accuracy.



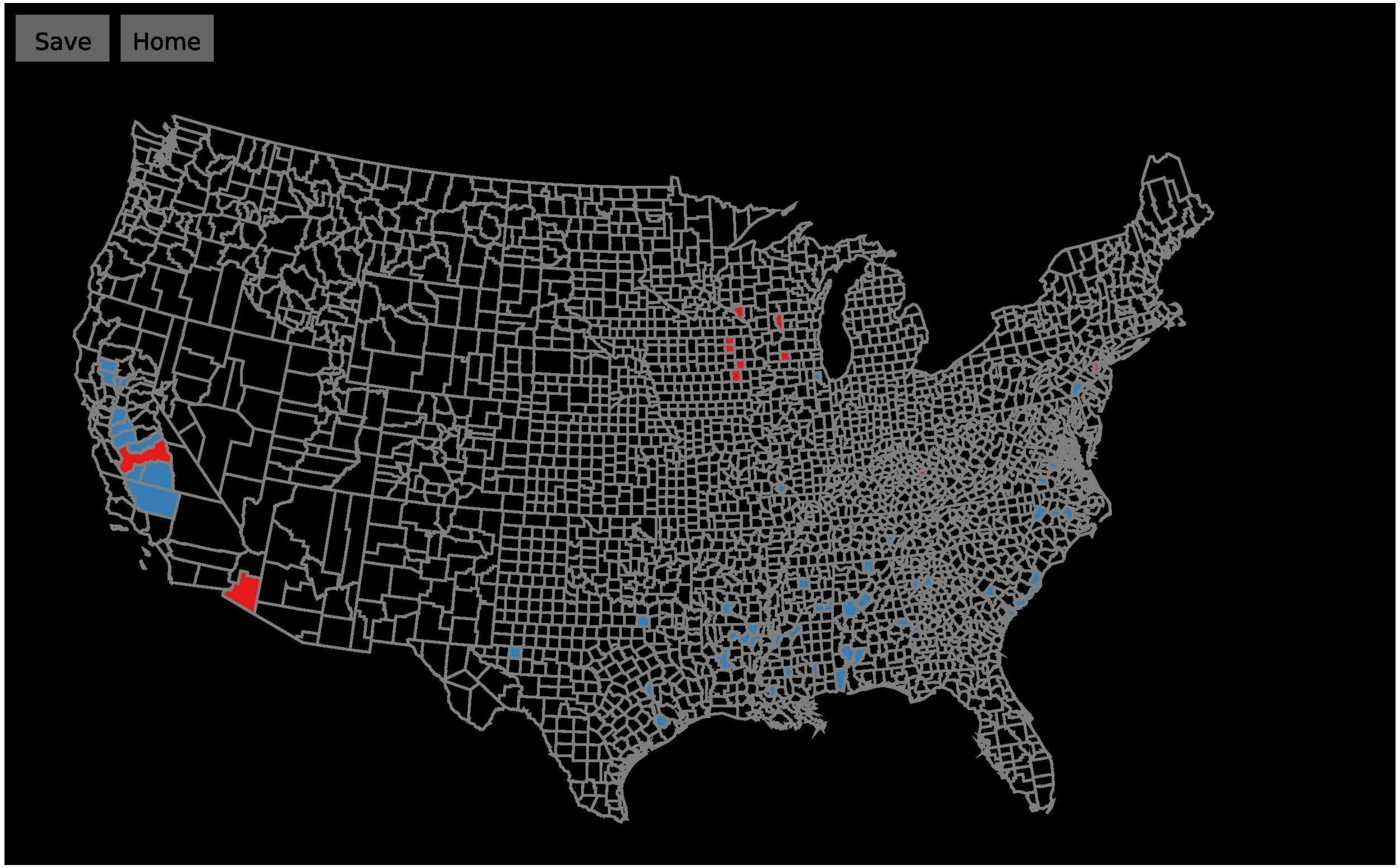
(a) Thematic map showing the spatial distribution of the 80 counties where the GWDA model resulted in the greatest reduction in classification uncertainty and coloured according to the actual US presidential election result (see figure 4.2(a)).

Figure 5.4: Visual analysis of 80 US counties where GWDA resulted in the greatest reduction in classification uncertainty compared to LDA.



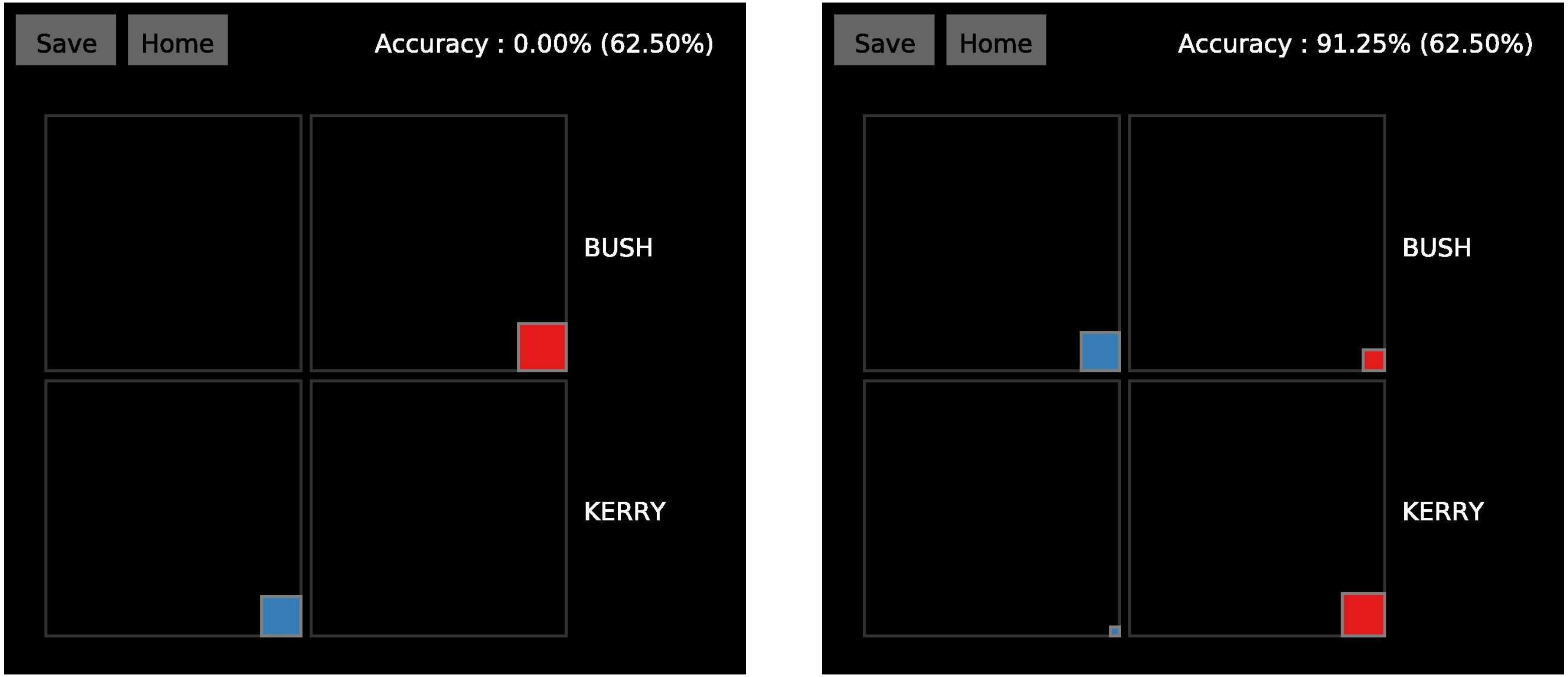
(b) Thematic map showing the spatial distribution of the 80 counties where the GWDA model resulted in the greatest reduction in classification uncertainty and coloured according to the LDA assigned US presidential election result (see figure 4.2(a)).

Figure 5.4: Visual analysis of 80 US counties where GWDA resulted in the greatest reduction in classification uncertainty compared to LDA.



(c) Thematic map showing the spatial distribution of the 80 counties where the GWDA model resulted in the greatest reduction in classification uncertainty and coloured according to the GWDA assigned US presidential election result (see figure 4.2(a)).

Figure 5.4: Visual analysis of 80 US counties where GWDA resulted in the greatest reduction in classification uncertainty compared to LDA.



(d) Fluctuation diagram showing the LDA confusion matrix for the 80 counties where the GWDA model resulted in the greatest reduction in classification uncertainty. Note that all 80 counties were misclassified by LDA.

(e) Fluctuation diagram showing the GWDA confusion matrix for the 80 counties where the GWDA model resulted in the greatest reduction in classification uncertainty. Note that only 7 counties were misclassified by GWDA.

Figure 5.4: Visual analysis of 80 US counties where GWDA resulted in the greatest reduction in classification uncertainty compared to LDA.

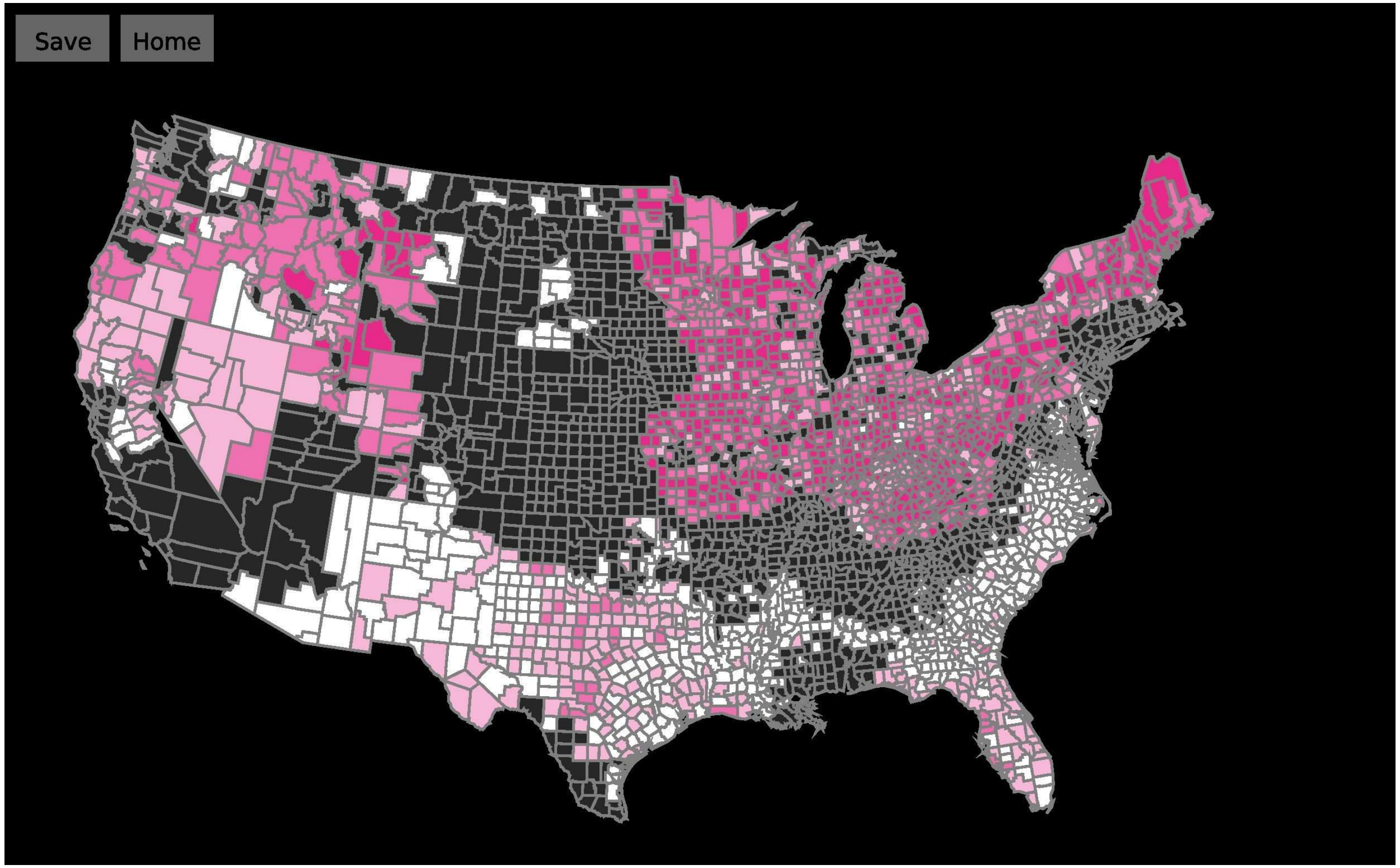
5.4.4 Exploring the extent of spatial non-stationarity

The techniques described in section 4.7 were used to explore the nature of spatial non-stationarity in the GWDA classification function parameters. Figure 4.10(b) shows a thematic map where counties are coloured by the most influential variable from the winning classification function. Spatial clustering is evident in this map and the link between the parameter values and the corresponding predictor variables was explored using a combination of the GWDA parameter legend (figure 4.10(a)), two thematic maps and the GWDA parameter treemap (figure 4.10(c). This works as follows:

One of the thematic maps was coloured according to the value of the *% white* variable by clicking on the colour button of the legend for this variable in figure 4.2(e). Next, the GWDA parameter treemap was sorted according to the value of the *% white* parameter (figure 4.10(c)). Then, all counties where the *% white* parameter is most important were selected by clicking on this category in the GWDA parameter legend (figure 4.10(a)). The results can be seen in figure 5.5. Note that all other counties are 'hidden' which permits further selections within this subset.

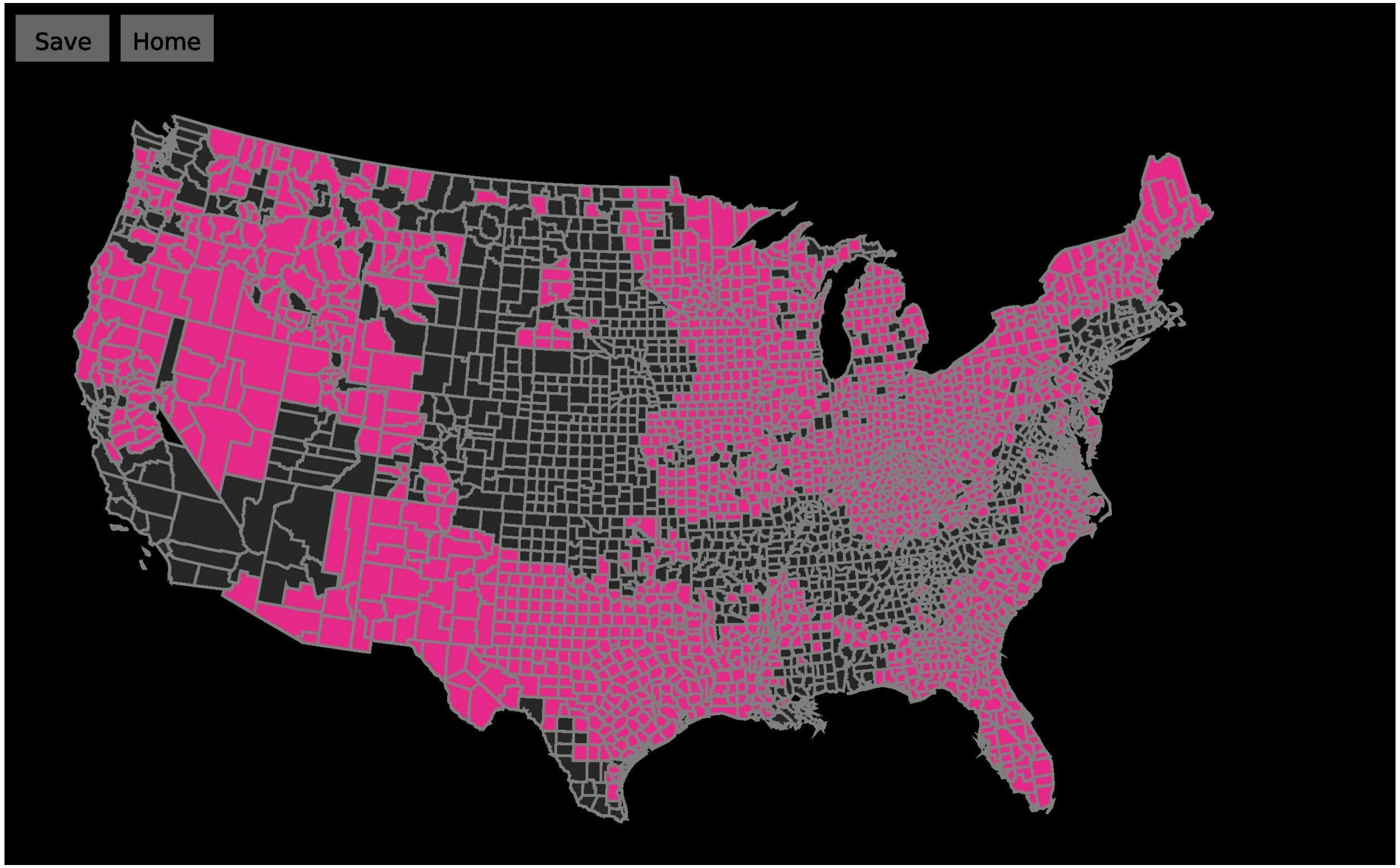
Figure 5.5(a) shows the spatial distribution of the *% white* variable for counties where the *% white* parameter was most important in the classification. Counties in the northern part of the US tend to have a higher percentage white than counties in the south. The GWDA parameter treemap in figure 5.5(c) is bimodal. The bottom half of the treemap shows counties where the GWDA parameter value for *% white* is most influential and negative (-8.19 to -0.47) and the top half shows counties where it most influential and positive (0.3 to 66.97). Selecting the all counties in the bottom half of the treemap shows the location of counties where this parameter is negative. These counties are located in the southern part of the US and are associated with low values of the *% white* parameter (figure 5.5(c)). Selecting the all counties in the top half of the treemap shows the location of counties where this parameter is positive. These counties are located in the north-eastern and north-western part of the US and are associated with high values of the *% white* parameter (figure 5.5(d)).

The conclusion is that counties where *% white* GWDA parameter is most influential and negative tend to be associated with low values of the *% white* variable. Conversely, counties where *% white* GWDA parameter is most influential and positive tends to be associated with higher values of the *% white* variable. Repeating the analysis for the other four predictor variables yielded a similar result.



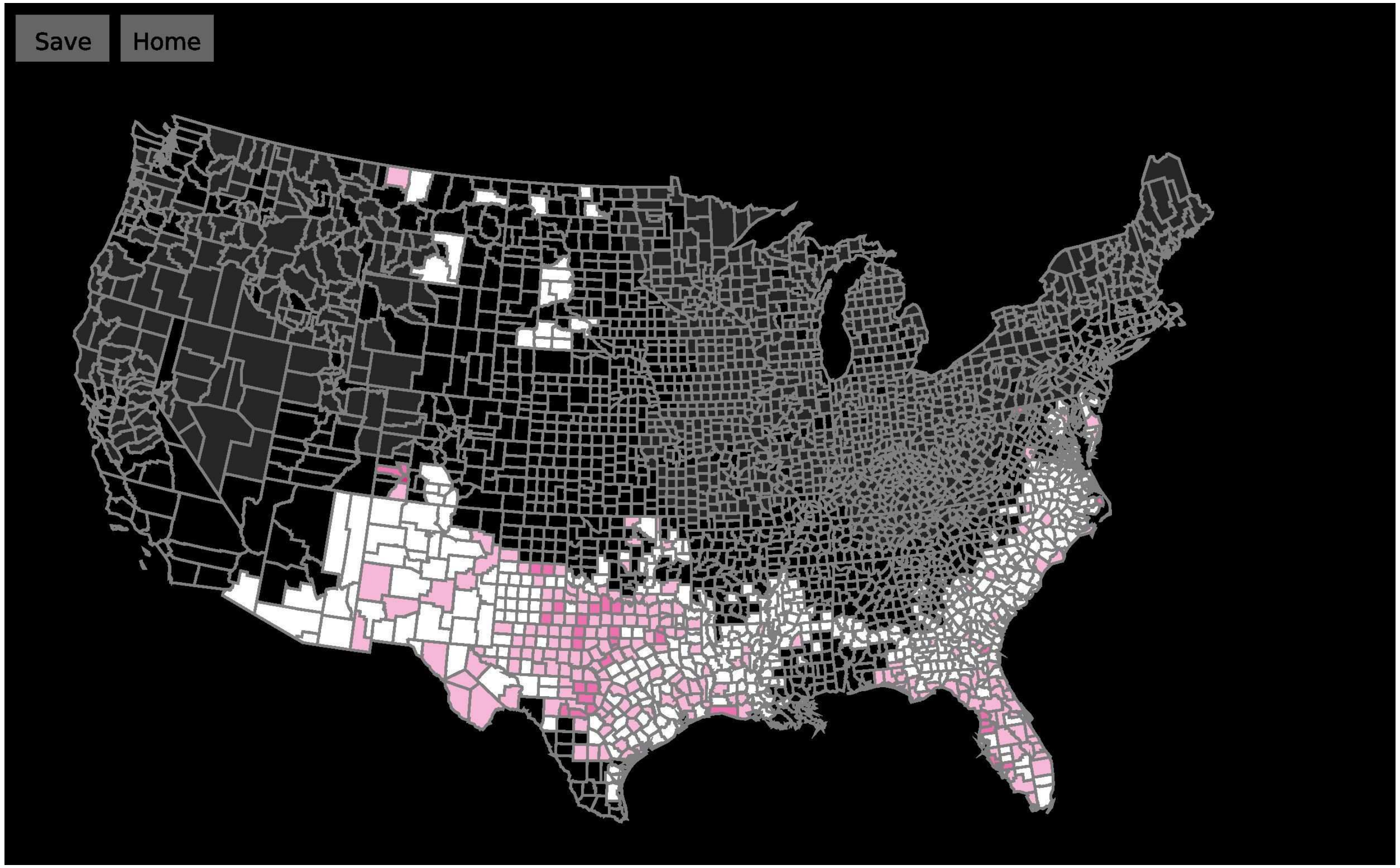
(a) Thematic map showing the spatial distribution of the 1,912 US counties where the most influential GWDA parameter is % *white* and coloured by the % *white* variable (see the legend in figure 4.2(e)).

Figure 5.5: Visual analysis of 1,912 US counties where the most influential GWDA parameter is % *white*.



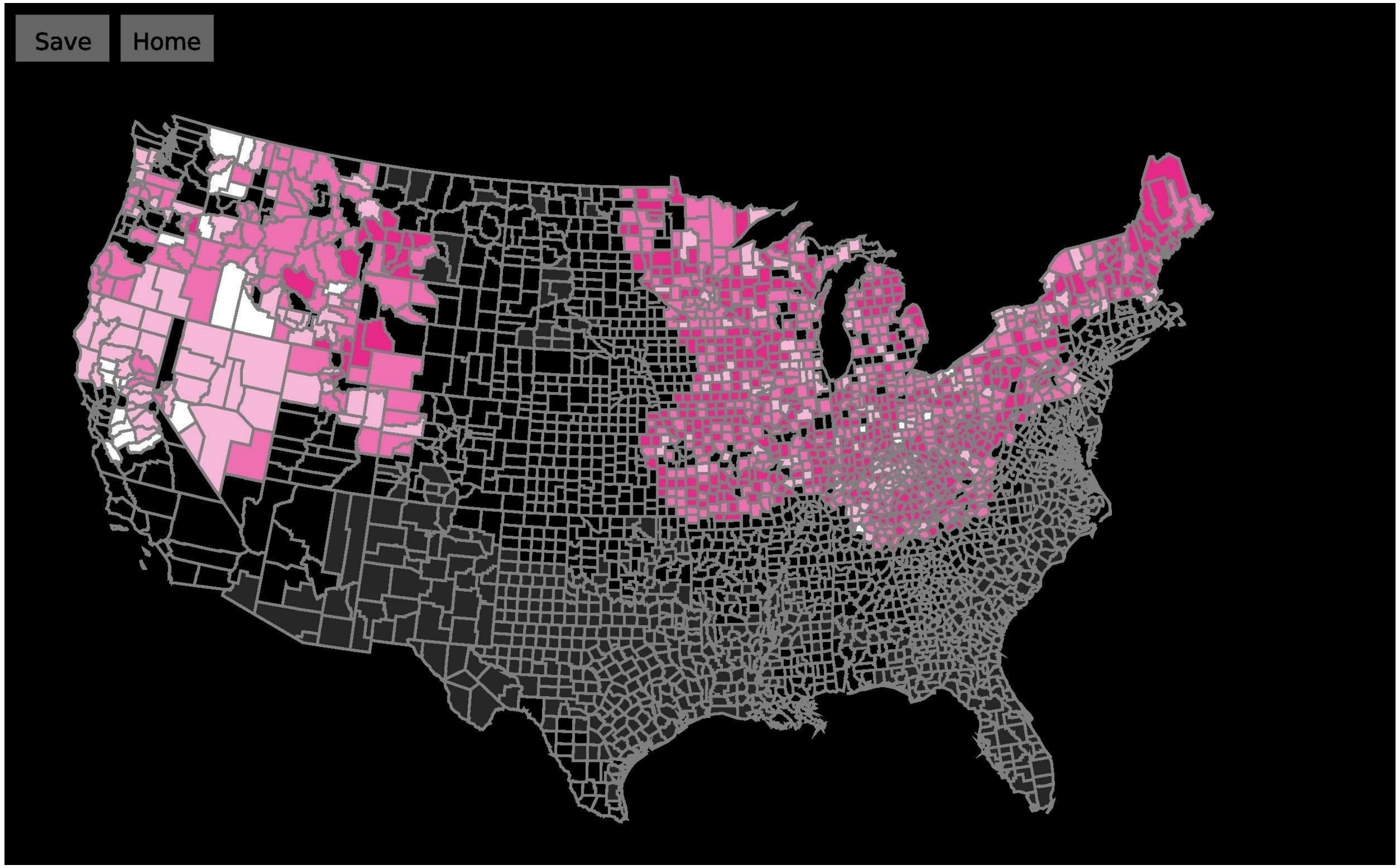
(b) Thematic map showing the spatial distribution of the $1,912$ US counties where the most influential GWDA parameter is *% white* and coloured by the *% white* parameter (see the legend in figure 4.10(a)).

Figure 5.5: Visual analysis of $1,912$ US counties where the most influential GWDA parameter is *% white*.



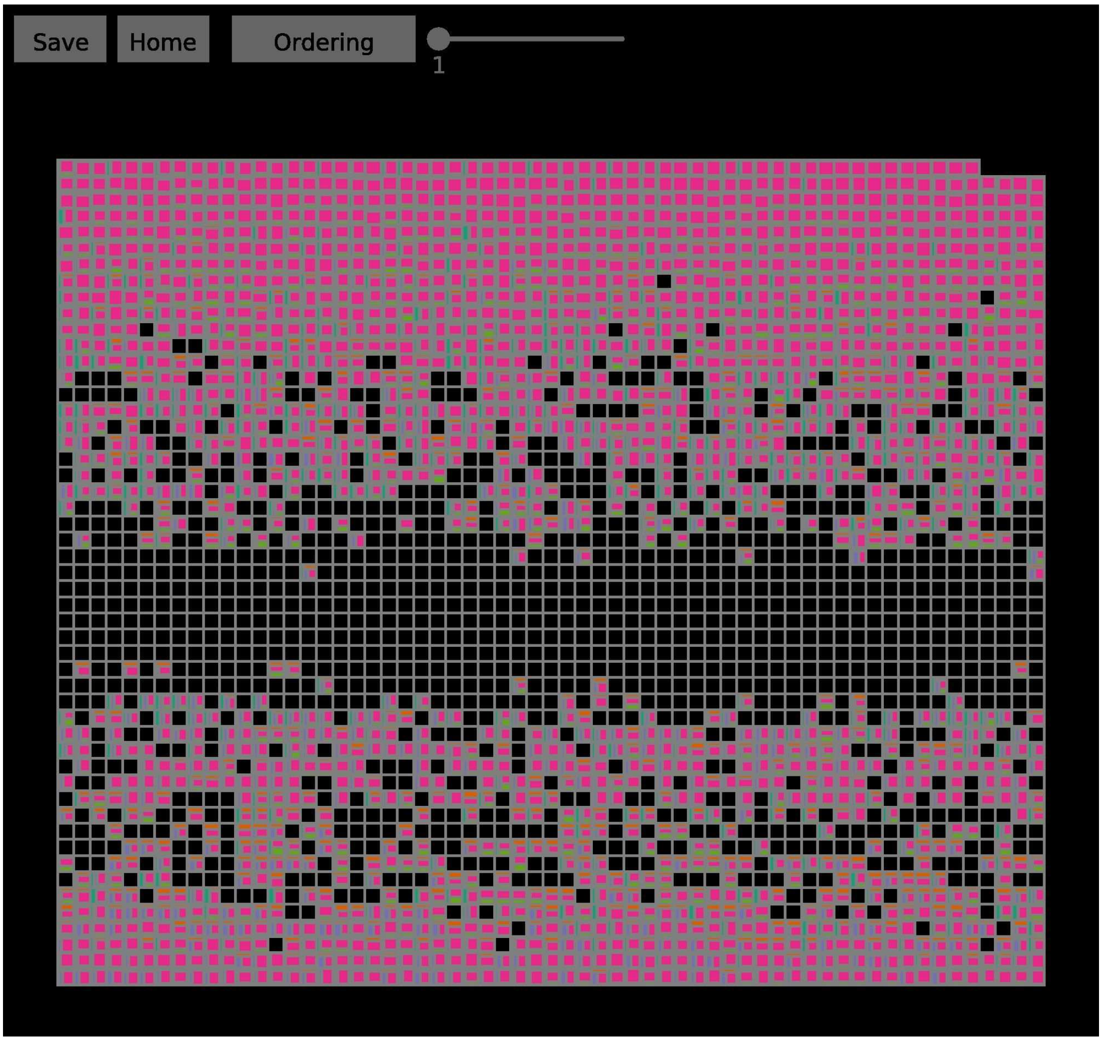
(c) Thematic map showing all counties from figure 5.5(a) where the GWDA parameter value for *% white* is negative.

Figure 5.5: Visual analysis of $1,912$ US counties where the most influential GWDA parameter is *% white*.



(d) Thematic map showing all counties from figure 5.5(a) where the GWDA parameter value for *% white* is positive.

Figure 5.5: Visual analysis of 1, 912 US counties where the most influential GWDA parameter is *% white*.



(e) Interactive treemap showing the bimodal distribution of the $1,912$ US counties where the most influential GWDA parameter is *% white*. The treemap cells are coloured according to the legend in figure 4.10(a).

Figure 5.5: Visual analysis of $1,912$ US counties where the most influential GWDA parameter is *% white*.

6 Conclusions & discussion

6.1 Conclusions

GWDA is used to model non-stationary relationships in categorical spatial data. It is a geographically local version of LDA that allows the relationship between a categorical dependent variable and a set of continuous predictor variables to vary spatially. This is referred to as spatial non-stationarity. On the contrary, LDA assumes that this relationship is fixed or stationary. If spatial non-stationarity is present in the data, the GWDA model of the relationship between the categories and the predictor variables should be superior to the LDA model.

The geovisual analytic tools and methodology described in this thesis enable an assessment of the performance of the GWDA model in terms of classification accuracy, classification uncertainty and spatial non-stationarity. The LDA model was used as a benchmark to assess the classification accuracy and classification uncertainty of the GWDA model. The classification accuracies of LDA and GWDA were investigated and compared using interactive fluctuation diagrams that visualised the confusion matrices together with two dynamically linked thematic maps. The GWDA classification uncertainty was explored using an interactive treemap visualisation of the posterior probabilities that was also dynamically linked with two thematic maps. The classification uncertainties associated with the GWDA and LDA models were compared using an odds ratio of the GWDA and LDA posterior probabilities for the actual categories that was dynamically linked with a thematic maps and two fluctuation diagrams. This fulfills the first research goal (see section 1.3). Finally, the extent and nature of spatial non-stationarity in the GWDA classification function parameters was explored using a treemap visualisation of the magnitude of the parameters associated with each classification function together with a thematic map showing the spatial distribution of the most significant parameter in terms of magnitude and an interactive legend. This fulfills the second research goal (see section 1.3).

The geovisual analytics methodology was demonstrated in a case study that used GWDA to model the relationship between the outcome of the 2004 US presidential election results and five socio-economic indicators.

6.2 Discussion

To date, the only previous attempt to visualise the output of GWDA is described in the paper by Brunsdon et al. (2007). The geovisual analytic tools described in chapter 4 build on their approach and improve it by adding an uncertainty analysis and enabling an exploration of non-stationarity in the classification functions for more than two predictor variables.

The results from the case study described in chapter 5 suggest a number of potential improvements and provide a direction for possible future developments, as follows.

Scalability to larger data sets is a significant limitation of the treemap visualisations described in sections 4.5 and 4.7. As the number of objects in the data set increases, the area available for each treemap cell will shrink, assuming that screen space and resolution remain constant. Even if the number of objects is held constant, increasing the number of categories and/or the number of predictor variables will result in further subdivisions of the fixed area treemap cells. This will make the treemap much harder to read. The use of a slider to control magnification of brushed objects is useful but the context provided by surrounding objects is still lost with larger data sets.

The use of area to represent magnitude in both the treemap and the fluctuation diagram can be difficult for users to interpret and compare. Lewandowsky and Spence (1989) recommend that length instead of area should be used to represent magnitude. The treemap layout algorithm used in this thesis is based on simple recursive splits of space, alternating between the horizontal and vertical (Urbanek 2008). A *slice and dice* treemap layout which splits the rectangles consistently along either the vertical or horizontal axis (Wood and Dykes 2008) may be an alternative. Using this approach, comparison of posterior probabilities or classification function parameters would be based on length along only one axis. The fluctuation diagram could also be modified so that magnitude is represented by the tile height rather than tile area. These changes could improve readability of both the treemap and the fluctuation diagram.

The recent work by Slingsby et al. (2011) is noted. They use interactive graphics to understand the nature of uncertainty in the UK output area classification. Their focus and methodology is different to the approach taken in this research but their ideas are related to this work. Their use of lightness to encode relative similarity to the category could be applied to this work to map confidence in the predicted categories. Additionally, summarizing the distribution profiles of each category in a PCP is easily understandable and scales much better than the techniques suggested here.

Another potential improvement would be to integrate additional descriptive statistics with the visualisations. The development of a statistical measure to quantify the degree of non-stationarity in the GWDA classification functions would be an example.

In section 2.5.4, seven different outputs from GWDA were listed. However, the geovisual analytics methodology described in chapter 4 only uses the first four of these. Incorporating the spatially varying bandwidth and the geographically weighted category means and covariance matrix into the geovisual analytics methodology could add additional insight into the relationship between the categories and the predictor variables. The spatially varying bandwidth provides useful information on the geographic scale of spatial variation. Visualising the spatial patterns for the geographically weighted category means and covariance matrix might prove useful in understanding the nature of spatial non-stationarity in the data. The exploratory visualisations of geographically weighted summary statistics by Dykes and Brunsdon (2007) are noted in this regard.

It was found that screen space is a significant limitation and merging of the visualisations could help address this. For example, Steed et al. (2009) successfully merges boxplots with a PCP.

A history function to capture user's actions would also be desirable. Jern (2009) has developed and implemented the concept of a 'story' to record the progress of users in the exploration. This could form a useful starting point.

Finally, in visual analytics, which includes the human aspect of visual data exploration, it is important to measure the ability of interactive visualisation systems to assist sense-making. This is challenging (Kang et al. 2011) and outside the scope of this thesis. However, the design and implementation of an evaluation experiment involving user testing would provide a more objective validation of the usability and utility of the tools than the case study.

References

Afifi, A., Clark, V. A., and May, S. (2004). *Computer-Aided Multivariate Statistics*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, Florida, fourth edition.

Ahmed, K. I., Demšar, U., and Monteys, X. (2009). Examining statistical segmentation of multibeam backscatter images with Geovisual Analytics. In *Proceedings of the ICA Workshop on Geospatial Analysis and Modelling*, pages 1–15.

Andrienko, G. and Andrienko, N. (2001). Constructing Parallel Coordinates Plot for Problem Solving. In *Proceedings of the Symposium on Smart Graphics*, pages 9–14.

Andrienko, G., Andrienko, N., Demšar, U., Dransch, D., Dykes, J., Fabrikant, S. I., Jern, M., Kraak, M.-J., Schumann, H., and Tominski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10):1577–1600.

Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M.-J., MacEachren, A. M., and Wrobel, S. (2007). Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, 21(8):839–857.

Brunsdon, C. (2009). Reply: GWDA and UK 2005 Election Results. *Geographical Analysis*, 41(3):338–341.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (2002a). Geographically Weighted Local Statistics Applied to Binary Data. In *Geographic Information Science*, volume 2478 of *Lecture Notes in Computer Science*, pages 38–50. Springer Verlag, Berlin / Heidelberg.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (2002b). Geographically weighted summary statistics — a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, 26(6):501–524.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (2007). Geographically Weighted Discriminant Analysis. *Geographical Analysis*, 39(4):376–396.

Crespo, R. (2009). *Statistical Extensions of GWR: Spatial Interpolation and a Spatiotemporal Approach*. PhD thesis, National University of Ireland, Maynooth.

Danese, M., Demšar, U., Masini, N., and Charlton, M. (2009). Investigating Material Decay of Historic Buildings using Visual Analytics with Multi-temporal Infrared Thermographic Data. *Archaeometry*, 52(3):482–501.

Demšar, U., Fotheringham, A. S., and Charlton, M. (2008a). Combining Geovisual Analytics with Spatial Statistics: the Example of Geographically Weighted Regression. *The Cartographic Journal*, 45(3):182–192.

Demšar, U., Fotheringham, A. S., and Charlton, M. (2008b). Exploring the spatio-temporal dynamics of geographical processes with geographically weighted regression and geovisual analytics. *Information Visualization*, 7(3-4):181–197.

Demšar, U. and Harris, P. (2011). Visual Comparison of Moving-Window Kriging models. *Cartographica*, 46(4):211–226.

Dykes, J. and Brunsdon, C. (2007). Geographically Weighted Visualization: Interactive Graphics for Scale-Varying Exploratory Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1161–1168.

Dykes, J., MacEachren, A. M., and Kraak, M.-J. (2005). *Exploring Geovisualization*, chapter 1. Exploring Geovisualization, pages 3–19. Elsevier, Oxford, UK.

Fotheringham, A. S., Brundson, C., and Charlton, M. (2002). *Geographically Weighted Regression — the analysis of spatially varying relationships.* John Wiley & Sons, Chichester, England.

Fotheringham, A. S. and Reeds, L. G. (1979). An Application of Discriminant Analysis to Agricultural Land Use Prediction. *Economic Geography*, 55(2):114–122.

Gahegan, M., Takatsuka, M., Wheeler, M., and Hardisty, F. (2002). Introducing GeoVISTA Studio: an integrated suite of visualization and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems*, 26(4):267–292.

Hardisty, F. and Robinson, A. C. (2011). The GeoViz Toolkit : Using component-oriented coordination methods for geographic visualization and analysis The GeoViz Toolkit : Using component-oriented coordination methods for geographic visualization and analysis. *International Journal of Geographical Information Science*, 25(2):191–210.

Harris, P., Brunsdon, C., and Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, 25(10):1717–1736.

Hofmann, H. (2008). *Handbook of Data Visualization*, chapter III.13 Mosaic Plots and Their Variants, pages 617–642. Springer Verlag, Berlin / Heidelberg.

Huang, B., Wu, B., and Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24(3):383–401.

Huisman, O., Santiago, I., Kraak, M.-J., and Retsios, B. (2009). Developing a Geovisual Analytics Environment for Investigating Archaeological Events: Extending the Space-Time Cube. *Cartography and Geographic Information Science*, 36(3):225–236.

Hurley, C. B. (2004). Clustering Visualizations of Multidimensional Data. *Journal of Computational and Graphical Statistics*, 13(4):788–806.

Inselberg, A. (2002). Visualization and data mining of high-dimensional data. *Chemometrics and Intelligent Laboratory Systems*, 60(1-2):147–159.

Jankowski, P., Andrienko, G., Andrienko, N., and Kisilevich, S. (2010). Discovering Landmark Preferences and Movement Patterns from Photo Postings. *Transactions in GIS*, 14(6):833–852.

Jern, M. (2009). Collaborative Web-Enabled GeoAnalytics Applied to OECD Regional Data. In *Cooperative Design, Visualization, and Engineering*, volume 5738 of *Lecture Notes in Computer Science*, pages 32–43. Springer Verlag, Berlin / Heidelberg.

Johnston, R. and Pattie, C. (2009). Comment: Geographically Weighted Discriminant Analysis and the 2005 British General Election. *Geographical Analysis*, 41(3):333–337.

Kang, Y.-A., Görg, C., and Stasko, J. (2011). How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):570–583.

Keim, D., Kohlhammer, J., Ellis, G., and Mansmann, F., editors (2010). *Mastering the Information Age: Solving Problems with Visual Analytics.* Eurographics Association. Available from : http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf.

Keim, D., Panse, C., and Sips, M. (2003). Visual Data Mining of Large Spatial Data Sets. In *Databases in Networked Information Systems*, volume 2822 of *Lecture Notes in Computer Science*, pages 201–215. Springer Verlag, Berlin / Heidelberg.

Keim, D., Panse, C., Sips, M., and North, S. C. (2004). Pixel based visual data mining of geo-spatial data. *Computers & Graphics*, 28(3):327–344.

Klecka, W. R. (1980). *Discriminant Analysis.* Quantitative Applications in the Social Sciences. Sage Publications, Beverley Hills, California.

Kraak, M.-J. (2008). From Geovisualisation Toward Geovisual Analytics (Editorial). *The Cartographic Journal*, 45(3):163–164.

Lewandowsky, S. and Spence, I. (1989). The Perception of Statistical Graphs. *Sociological Methods & Research*, 18(2-3):200–242.

Luo, D., Yang, J., Krstajic, M., Ribarsky, W., and Keim, D. (2012). EventRiver: Visually Exploring Text Collections with Temporal References. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105.

MacEachren, A. M. and Kraak, M.-J. (2001). Research Challenges in Geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12.

Manly, B. F. J. (2005). *Multivariate Statistical Methods : A Primer.* Chapman & Hall/CRC, Boca Raton, Florida, third edition.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley & Sons, Hoboken, New Jersey.

Reimann, C., Filzmoser, P., Garrett, R., and Dutter, R. (2008). *Statistical Data Analysis Explained: Applied Environmental Statistics with R.* John Wiley & Sons, Chichester, England.

Rogerson, P. A. (2008). *Statistical Methods for Geography.* Sage Publications, London, second edition.

Serra, P., More, G., and Pons, X. (2007). Monitoring winter flooding of rice fields on the coastal wetland of Ebre delta with multitemporal remote sensing images. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 2495–2498.

Sharma, S. (1996). *Applied Multivariate Techniques.* John Wiley & Sons, New York.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343.

Shneiderman, B. (2002). Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization*, 1(1):5–12.

Slingsby, A., Dykes, J., and Wood, J. (2011). Exploring Uncertainty in Geodemographics with Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 17(6).

Slocum, T. A., McMaster, R. B., Kessler, F. C., and Howard, H. H. (2009). *Thematic Cartography and Geovisualization*. Prentice Hall Series in Geographic Information Systems. Pearson Prentice Hall, New Jersey, USA, third edition.

Steed, C., Fitzpatrick, P., Swan, J., and Jankun-Kelly, T. (2009). Tropical Cyclone Trend Analysis Using Enhanced Parallel Coordinates and Statistical Analytics. *Cartography and Geographic Information Science*, 36(3):251–265.

Theus, M. (2008). *Handbook of Data Visualization*, chapter II.6 High-dimensional Data Visualization, pages 151–178. Springer Verlag, Berlin / Heidelberg.

Theus, M. and Urbanek, S. (2009). *Interactive Graphics for Data Analysis: Principles and Examples*. Chapman & Hall/CRC, Boca Raton, Florida.

Thomas, J. J. and Cook, K. A., editors (2005). *Illuminating the Path: The Research Agenda for Visual Analytics*. National Visualization and Analytics Center. Available from: http://nvac.pnl.gov/docs/RD_Agenda_VisualAnalytics.pdf.

Tomaszewski, B., Blanford, J., Ross, K., Pezanowski, S., and MacEachren, A. M. (2011). Supporting geographically-aware web document foraging and sensemaking. *Computers, Environment and Urban Systems*, 35(3):192–207.

Urbanek, S. (2008). *Handbook of Data Visualization*, chapter II.10 Visualizing Trees and Forests, pages 243–264. Springer Verlag, Berlin / Heidelberg.

Wegman, E. J. (1990). Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*, 85(411):664–675.

Wills, G. (2008). *Handbook of Data Visualization*, chapter II.9 Linked Data Views, pages 217–241. Springer Verlag, Berlin / Heidelberg.

Wilmut, M., Bloomer, S., and Preston, J. (2009). Discriminant Analysis in Image-Based Seabed Classification. In *Proceedings of the Underwater Acoustics Measurements: Technologies & Results Conference.*

Wood, J. and Dykes, J. (2008). Spatially Ordered Treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1348–55.

Appendices

A short movie demonstrating the geovisual analytics software described in Chapters 4 and 5 can be seen at `http://vimeo.com/28954460`.

This software has been developed using the Processing 1.5.1 Java Libraries (`http://processing.org/`). The following Java Libararies are also used:

- giCentreUtils 3.1 (`http://gicentre.org/utils/`).
- Apache Commons Math 2.1 (`http://commons.apache.org/math/`).
- GeoTools 2.7 (`http://geotools.org/`).
- Java Topology Suite 1.11 (`http://www.vividsolutions.com/jts/`).

Colours were specified using the ColorBrewer 2 web application (`http://colorbrewer2.org`, accessed on Wednesday 8th February, 2012).